Rialtas na hÉireann
Government of Ireland

Digital Plan for the Irish Language

Speech and Language Technologies 2023-2027

Prepared by the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media
www.gov.ie

Table of Contents

Introduction

Part I: Resources

Part II: Core Technologies

Part III: Applications

Conclusions

Message from the Minister and Minister of State

We live today in a 'Digital Age' where technology is an integral part of the way we live our lives. Ireland has been a major European hub for modern technology for some time, and while English is the dominant in this area in general, Irish has been in a much weaker position up to now. Our national language is one of Ireland's most valuable treasures in terms of culture and heritage and is a manifestation of our unique identity as a nation. Now is the time for us to focus on integrating this heritage with today's digital society.

This Plan is a roadmap for realising our vision in the years ahead. Our objective is to enable the Irish-language community to build and revitalise the language by establishing and developing a vibrant Irish-language technology sector, enabling this community to take up new opportunities to participate fully in modern day society. Technologies in the field of speech and language, in particular, need to be developed, so that our language keeps pace with major European languages in terms of application and usage of those aspects of modern technology. Success in this regard will greatly contribute to the normalisation of the use of Irish in society as a whole. Indeed, we envisage that the language will flourish in the digital age if the opportunities provided by new technology are fully tapped.

The Digital Plan is a national plan that aligns with the priorities of the European Union in terms of the digital transition. This Plan also aligns with the 20 Year Strategy for the Irish Language 2010-2030, in particular Area 5 of the Strategy: Media and Technology. This document has been prepared also in the context of other related policies and legislation, including AI - Here for Good: National Artificial Intelligence Strategy for Ireland; Harnessing Digital - The Digital Ireland Framework; Action Plan for the Irish Language, 2018-2022; and the Disability Act 2005.

This Digital Plan provides an overview of the research required to make Irish-language linguistic resources available in the coming years. The underlying speech and language technologies related to these linguistic resources are also set out as well as the aids and applications which need to be developed into the future. People with disabilities, a vulnerable group, will be a particular focus. We have set out to enable anyone interested in the language to have inclusive access to support and facilities in order to easily use Irish in the digital space on a daily basis. Accordingly, the Plan supports state-of-the-art technologies and applications in a variety of areas: multi-dialect speech technology, enhanced search engines, machine translation, educational platforms and apps that will assist with the teaching of the language as well as screen readers, speech synthesis and other aids for visually impaired Irish speakers, to name but a few.
Implementation of this Plan will require cross-departmental support and a parallel integrated cross-disciplinary effort. Linguists, language engineers and other experts will have a part to play in carrying out the pioneering work that is envisaged. Our Department's officials will work to explore all the possibilities for implementation of the Plan, including the opportunities for the establishment of new structures, through which the technologies developed in the various bodies will be overseen and led on an integrated basis. As the new technologies for Irish come to fruition, we expect them to be used in private sector businesses to provide an Irish language service to end users. The Irish-language community will play an active role in all stages of the implementation of the Plan so that the research is guided by the needs of users.

Irish is unique as a minority language because of its constitutional status as the State's national language and its recognition as an official language of the European Union. Our view is, therefore, that this Plan will serve as a model for the future development and strengthening of other minority languages.

It is widely understood that technology is in a state of constant change and evolution. As the research envisaged here develops, and as results and applications derive from that research, the Plan will be formally reviewed and updated once every five years. In addition to the formal reviews, we will maintain a watchful eye on developments in this area, with priorities being revised as necessary.

We look forward to working with our stakeholders and the Irish language community as a whole in the coming years to deliver on the ambitious vision and goal that we have set out in this Plan to promote the growth and development of the language in this vital area of our lives.

Catherine Martin TD
Minister for Tourism, Culture, Arts, Gaeltacht,
Sport and Media

Jack Chambers TD
Government Chief Whip and Minister of State
for the Gaeltacht and Sport

December 2022

Acknowledgements

This Plan was developed through the excellent work carried out by a number of talented researchers commissioned by the Department. We are particularly grateful to the authors: Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Elaine Uí Dhonnchadha at Trinity College, Dublin; Teresa Lynn and John Judge at Dublin City University.

During this work, the authors received advice on aspects of the technology and feedback on drafts of the Plan from other experts in Ireland and abroad. We are very grateful to them all, and especially: Brian Ó Raghallaigh, Mícheál Ó Meachair, Brian Davis, Monica Ward at Dublin City University; Emily Barnes, Andy Murphy, Harald Berthelsen, Liam Lonergan, Christer Gobl and Connor McCabe at Trinity College, Dublin; Kevin Scannell at St. Louis University in the United States; Mengjie Qian at the University of Cambridge; Peter Bell, Simon King, Steve Renals and Christian Veaux at the Centre for Speech Technology Research, University of Edinburgh; Martin Russell at the University of Birmingham; Shane Ó hEidhin, Micheál Ó Maolruanaidh at Foras na Gaeilge.

Executive Summary

Speech and language technologies permeate every aspect of modern life. The digital revolution is English-language dominated and is described as a 'digital timebomb' for minority and endangered languages, accelerating the catastrophic rate at which many of the world's languages are being lost. Language transmission depends on the younger, tech-savvy generation, and technology provision for one's own language will undoubtedly be a critical factor in its On the positive side, these technologies and the linguistic knowledge that underpins them offer unprecedented opportunities for language maintenance and revitalisation.

This document outlines the research and development needed to bring Irish into the digital age. It will require an interdisciplinary effort, involving linguists, engineers, computational linguists, speech scientists, and software specialists. It is important that this work is carried out in close collaboration with the Irish language community, and with stakeholders in areas such as education, disability and access.

The development priorities for Irish are not identical to those for languages like English, where the primary goal is commercial. In the case of Irish, it is critical that development focuses on the requirements of the Irish language community and the goal of language maintenance. It is thus a guiding principle of the Digital Plan that the challenges addressed, and solutions adopted, should be tailored to the Irish context, and founded on a deep knowledge of the language. .

The research and development necessary to progress the Digital Plan is outlined in three strands of activity, presented in the three parts of this document. These are **Part I: Resources, Part II: Core Technologies**, and **Part III: Applications**. These three strands cover the range from basic to applied research and need to be progressed together.

For many readers, Parts II and III, which deal with the core technologies and real-world applications respectively will be the most accessible, and may be the best initial focus of interest. Part I contains more detailed, technical information concerning the basic resources that underpin the Plan. This will be of interest to those who seek a fuller understanding of this research area, and it will serve to clarify various aspects of the technology developments which are discussed in the later Parts II and III.

Chapters 2-15 deal with specific resources, technologies and applications. Considerable explanatory information is included in each case, and discussion of its importance and potential uses. The hope is that a better understanding of the field will stimulate the readers' interest and may inspire them to support and become part of this new and rapidly growing field. As the document is quite detailed, it has been laid out in such a way as to enable someone to get a quick overview of the field and to dip in and out of specific aspects that are of interest. For this reason, where possible, the chapters have a common format, covering for each topic the following questions, going from the simpler, more general, to the more detailed discussion.

1. What is it?
2. Why important and for whom?
3. How does it work?
4. What has been done to date?
5. Future work
6. Recommendations

References are included at the end of each chapter.

In the Introduction and Conclusion (Chapters 1 and 16), consideration is given to the broader picture, and to the actions that are needed to ensure the Digital Plan realises its potential impact for the language and for the language community. As mentioned, development must be sensitive to the Irish language context, and priority given to applications needed by the community, applications that can transform Irish language teaching, and applications that ensure the inclusion of all with disabilities.

The Plan should not be viewed as a finite task of providing items of technology, but as a long-term, continuous investment that will require:

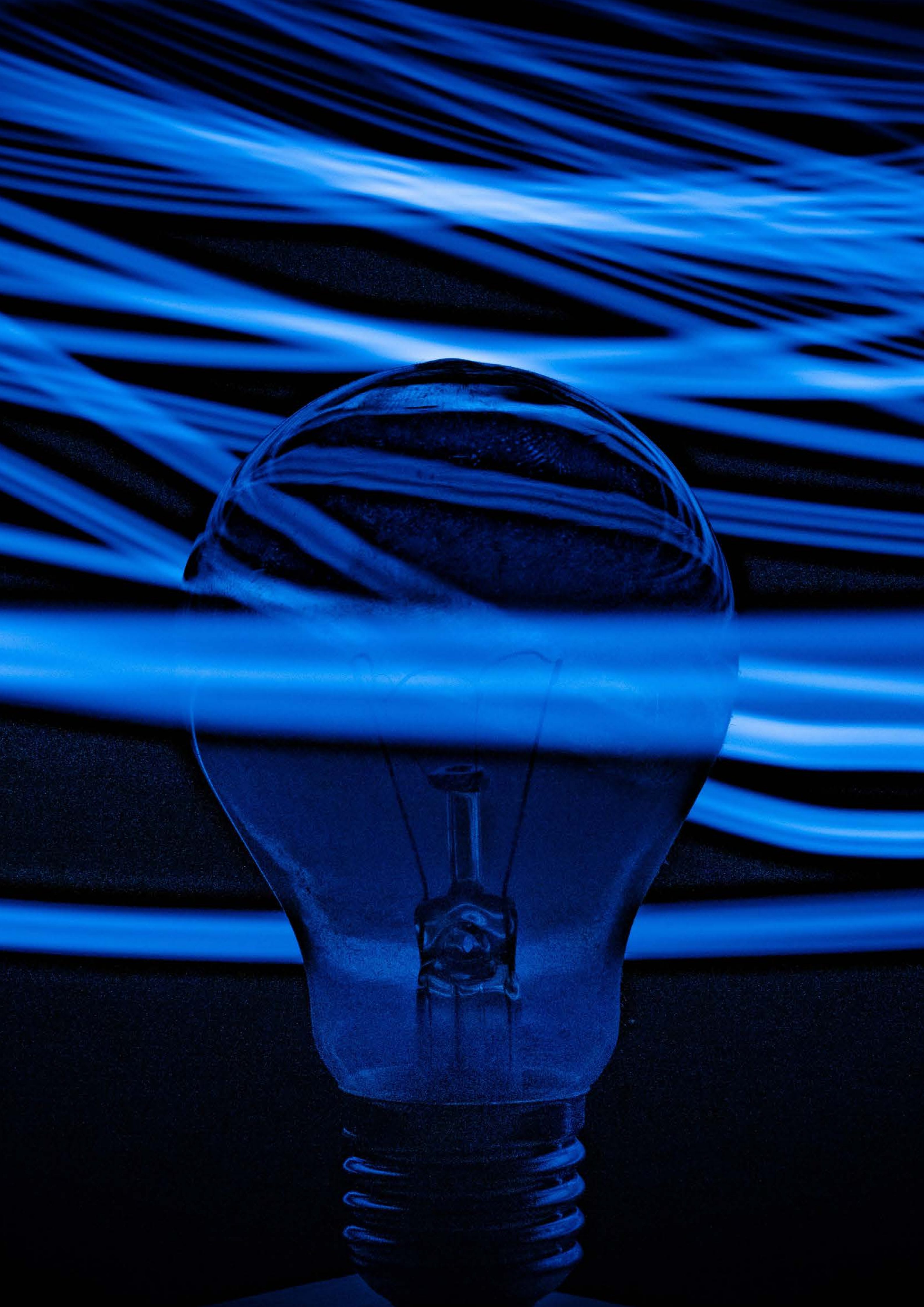
- **People: experts in the field:** a top priority is the education of the skilled researchers that will enable continuous future development of the technologies and their linguistic foundations – researchers with high levels of competence in Irish that can work in the interdisciplinary environment mentioned above.

- **Centres of Excellence:** should provide the physical and technical infrastructure that hosts Irish-language focussed interdisciplinary teams. These teams will provide the linguistic and technology advances that deliver state-of-the-art speech and language technology for Irish.
- **Digital Innovation Hubs:** downstream, as the technology and applications emerge, opportunities for further innovation and commercial ventures will follow. Forward planning is needed here, for example how campus companies might be established and clustered in Digital Hubs. The Gaeltacht would be an ideal home for such initiatives, though enterprises in other locations should also be encouraged. Close linkage between company/hub and the research centre would be important to ensure that the former are up to speed in this rapidly evolving field. Policy guidelines should be developed, and it is recommended that profits emerging from the research and innovation of the Digital Plan should feed back into supporting its longer-term research and development.
- **Community Engagement:** The Digital Plan will only succeed to the extent that *Pobal na Gaeilge*, engages with it – in the Gaeltacht, throughout the country, learners and proficient speakers alike, including those who through disability were excluded in the past, and including the many across the world who engage at any level with the language. By its nature, digital speech and language technology has the potential to link the diverse communities that comprise *Pobal na Gaeilge*, empowering them to realise through their combined energy the potential of the digital age for the Irish language. Towards this, we need to go beyond the typical research-outreach activities: we need to nurture community engagement with this initiative from grassroots level upwards, as active partners in numerous aspects of the research and development and in the dissemination of outputs.
- **Gaeltacht involvement:** The native speaker Gaeltacht community is particularly critical in the realisation of the Digital Plan and will, it is hoped, be centrally involved in the future multidisciplinary research that leads it. Suggested mechanisms for involving the young Gaeltacht native speakers would include co-location of educational and downstream commercial initiatives between (e.g. university-based) research centres and the Gaeltacht.
- **Regular Review:** technology is advancing rapidly and therefore ongoing review will be required, with adjustment to goals, to take account of the evolving technology and the situation on the ground. Formal 5-yearly reviews are recommended.
- **Funding:** Delivering on the Digital Plan will require large scale, continuous, and long-term investment, along with a commitment across Government Departments to ensure that the outputs of research and development will be available to all sectors of the Irish language community, and that they are continuously maintained and supported. In this endeavour, support and funding is to be expected from the major companies working in this sector in Ireland, as part of their corporate social responsibility – support that will enable the development of an indigenous, local Irish speech and language technology sector, firmly rooted in the Irish language community and drawing on a deep understanding of the language.

In terms of the world's endangered languages, Irish has been uniquely fortunate in the level of recognition and State support it enjoys. The intention is that the Digital Plan for Irish will serve as a model for other minority and lesser spoken languages, and that as the Plan evolves, our knowledge, experience and, where appropriate, our resources will be shared with other communities who struggle to maintain their language in the shifting landscape of the digital age.

With a proper level of investment and with effective management, the Digital Plan will greatly support the Irish language community and further the preservation of the language, weaving our linguistic heritage into the new technological era.

Introduction



Chapter 1
Irish in the Digital Age

1.1 Motivation and aims

Digital communication technology plays an increasingly central role in everyday life, in how we work, socialise, learn, connect to information, seek entertainment, control our environment etc. Speech and language technologies allow us to interact with computers and digital resources, either through text, or increasingly, through speech.

The digital revolution, described as a "digital timebomb" facing lesser-used language speakers [1], is feeding the globalisation of Anglocentric culture, fuelling the processes whereby minority languages are being lost at an unprecedented rate. However, this same digital technology offers unique opportunities to counteract this trend of language attrition to help document, preserve, maintain and revitalise endangered languages and dialects (see for example [2]). Precisely because it pervades every aspect of our lives, embedding Irish in technology facilitates its use in multiple domains of activity and harnesses the opportunities for the language to be part of this new digital age.

The Digital Plan presents a vision for how digital speech and language technologies may be developed for the Irish language. It provides an overview of goals for research and development, reviewing briefly what has been done towards these goals to date. The emphasis is on the living language, and on state-of-the-art technologies that will promote the successful acquisition and widespread communicative use of the language in all spheres. It is a basic principle of the Plan that all tools and resources developed will be made freely available to the community.

The intention is to provide high quality technologies and resources, on a par with what is available for the major languages. The Plan therefore complements Ireland's 2021 AI Strategy *"AI- Here for Good"*, by demonstrating the requirement to broaden such technologies beyond English and encompass support for the Irish language. As technology is ever-evolving, continuous long-term development of this research will be required, and the goals specified now will need to be regularly updated. Given the pace of change in this field, this plan will need to be reviewed and revised every five years.

This document presents a detailed overview of the areas that will require basic research and development. It situates this technology – the need for it and the potential it offers – in the Irish context, detailing how it can support the language and the Irish-language community. Given this focus, considerable explanatory information on these technologies and their relevance is included, in the hope that it will inform and inspire those who wish to support and become part of this enterprise. Although consequently quite long and technically detailed in parts, it is for the most part written in such a way as to be accessible to the non-specialist reader, aiming to provide an understanding of the field, of the Plan's goals and of the challenges to be met.

The specific areas of application that will be most important for Irish will often differ from those prioritised for major world languages such as English, where development is primarily driven by commercial considerations. Priorities for Irish must answer to the needs of the language community and the goals of language maintenance.

The research and development envisaged is presented in three parts. Part I deals with the provision of linguistic resources and tools; Part II covers the development of core technologies that deploy these resources; Part III concerns the development of user applications that

[1] There have been rapid technological advancements even since the drafting of this document began.

draw on the outputs of Parts I and II. Note that it is through these applications that the potential for major impact arises.

This Introduction, and the Conclusions (Chapter 16) look at the big picture as the Digital Plan involves much more than the supply of specific technologies. The goal is to establish a strong Irish-language speech and language sector in Ireland, which will continue to deliver state-of-the-art technologies, applications and resources into the future — and that are informed by a deep knowledge of the language structure, are developed with the participation of the language community, meet the community's needs and are sensitive to the sociolinguistic context and the specific challenges facing Irish today.

1.2 The digital age

Speech and language technologies are ubiquitous in the modern world. Social media is generating new communities and is pivotal to how young people socialise and interact: machine translation is creating new possibilities for communication across language barriers; we seek information from our virtual assistants on our phones; call centres are now typically operated using interactive voice (response) technology; sophisticated interactive games with speech input/output are increasingly being used in educational settings; mobile app technologies and online learning sites are making education more accessible to learners. These technologies are enabling communication for those who were excluded due to disability; increasingly we interact with a range of 'agents' that accrue data on our voice and behaviour in order to monitor and advise on our physical and mental health; robots and virtual reality characters can be 'companions' that help the elderly remain independent in their homes. Speech and language technology is everywhere, and in Ireland it is all happening through English.

1.3 Urgency of the task

The Digital Plan is urgently needed if Irish is to benefit from the positive aspects of the digital revolution and to survive the threat of digital extinction [3]. The recent *Report on the Irish Language* [4] from the 2022 *European Language Equality Project* [5] highlights Irish as being one of only two official EU languages falling under the category of "weak or no support" with respect to language technology. As elaborated in this document, although rapid development is expected in certain areas, sophisticated technologies and truly useful applications will not appear overnight. Rather, a broad-ranging programme of development is envisaged, which includes linguistic resources to underpin the building of core technology and applications. Such a programme requires infrastructural investment to provide for the education of personnel with the necessary language and technical skillsets. It also requires mechanisms to ensure that the emerging technologies are applied in ways that have an impact for the Irish language community. If these opportunities are not grasped now, the technology gap between Irish and English will continue to widen, with accelerated shrinkage of domains where we can use Irish in our everyday activities.

1.4 Challenges of the linguistic and sociolinguistic context

The linguistic and sociolinguistic landscape of Irish presents many challenges and imposes constraints which need to be foregrounded in setting priorities for the Digital Plan. Irish is an endangered language, spoken as an everyday community language by a relatively small population of native speakers in the Gaeltacht areas. Furthermore, the Gaeltacht communities are geographically scattered, so that the critical density of Irish language native users is further diluted. Even in these small Gaeltacht communities, English is gaining ground as the common language, particularly among the digitally-connected younger generation [6]. Outside the Gaeltacht, and particularly in urban centres, there are many families who use Irish as the language of the home, and many individuals who strive to become more proficient in it as a second working language in their lives – but again, these families/individuals are dispersed in the wider population.

On the positive side, Irish enjoys Government and Constitutional support as the first national language, and the *20 Year Strategy for the Irish Language 2010 – 2030* recognises the State's commitment to its revival [7]. *The Action Plan for the Irish Language, 2018-2022* [8] provided a more cohesive and coherent framework in support of the Strategy's implementation, which focused on specific, measurable, achievable, realistic and time-specific actions to be implemented over this period. It also included two specific areas, *Area 5: Media and Technology*, and *Area 6: Dictionaries*, which relate strongly to measures outlined in this Plan. Building on this, the present document greatly extends the remit of the envisaged research and development and emphasises the long-term vision of establishing a strong, local, community-led Irish speech and language technology sector.

Since the foundation of the State, it has been recognised that the education system would play a critical role in the transmission of Irish. It is a core curriculum subject until school leaving age. The teaching of Irish presents considerable educational challenges, compounded by the fact that a great many learners have no regular or easy access to native speakers of the language. Research has demonstrated a worrying decline in the language proficiency of pupils in English-medium schools [9]. At the same time, the striking grassroots-led growth in Irish-medium schools demonstrates a public appetite for Irish-language education. Abroad, the growth in Irish-language courses at University level and the widespread uptake of current web-based resources for Irish reveals an international community of Irish-language users and learners. This international dimension can strengthen and support local Irish language communities, and the emerging technologies will enable a more connected Irish-language community.

There are specific linguistic factors that impact the development of this Plan. The linguistic structure of Irish differs in major ways from English and other Western European languages (see Chapter 2) with implications for how we develop language technologies. Current technologies have been primarily developed for English, drawing on many decades of basic linguistic research on the language and they are therefore inevitably optimised to deal with the linguistic features of English. Empirical, technology-ready descriptions of the linguistic structure of Irish are needed, to ensure that existing technologies can be optimised to 'fit' the language and that this knowledge can be incorporated in applications that users of Irish need.

A further factor with implications for speech technologies is the fact that there is no single spoken standard variety in Irish, but rather three main dialects and a number of sub-dialects. A written standardisation, *An Caighdeán Oifigiúil* [10] draws on different dialects but does not represent the speech of any spoken variety. This effectively means that a multi-dialect approach is required for Irish speech technologies, such as speech synthesis, speech recognition and dialogue systems, as well as for applications that make use of them. For the major languages, technology development has focused on the single, standard spoken variety. The need to cater for multiple dialects of Irish from the outset presents a challenge – but is essential to ensure that the technologies developed are truly useful for the entire language community.

1.5 Plan outline

This plan is divided into three Parts, illustrated using the image of a tree in Figure 1.1. It is important to appreciate that developing speech and language technologies for Irish is not just about taking technologies and adapting them around the edges for Irish, but rather about a long-term process involving parallel basic and applied research. Like the tree, the fruits (applications) require extensive roots (linguistic resources) and a strong trunk (core technologies). Although presented as a logical sequence, research in all three areas needs to proceed in parallel and collaboratively to ensure that the emerging technology is appropriate for users and achieves the greatest possible impact.

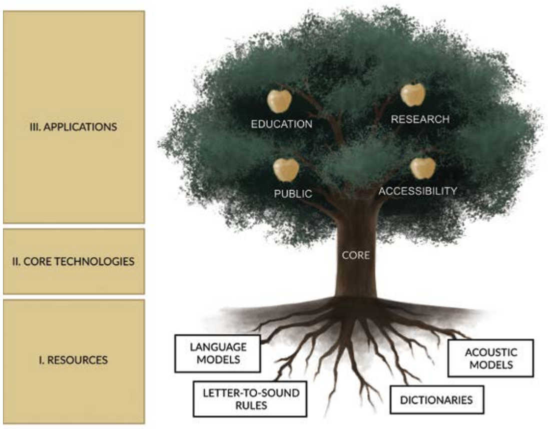


Figure 1.1: The three parts of the Digital Plan

Part I provides the foundations – the prerequisite **speech and language resources**, tools and models – the *roots* in Figure 1.1 These include digital corpora of spoken and written Irish, extensive digital linguistic analyses to provide the knowledge base, as well as specific components, tools and models needed for technology development (these are already highly developed for English through decades of basic research). This research will deepen our understanding of Irish and the resources are essential to many of the applications needed for Irish, especially for teaching and learning.

Part II concerns **core speech and language technologies**, the *tree trunk* that draw on the resources of Part I. These complex systems include text-to-speech synthesis, speech recognition, spoken dialogue systems, machine translation and information retrieval systems.

Part III concerns the **applications**, the *fruits*, where both the core technologies and the resources are harnessed for specific goals targeting the wider public as well as specific end-users. It is through these applications that the full potential impact will be felt by the public. The focus here is specifically on applications for broad public use, for education and for the inclusion of those with disabilities.

Note to the reader: For many readers, Part II (Core technologies) and Part III (Applications) are the most accessible, providing concrete examples that explain the technologies and illustrate their advantages. Part I (Resources) is more technical in nature and describes the fundamental linguistic and other research that underpins the technology. While one doesn't need to have read Part I to make sense of Parts II & III, it nevertheless provides the background for the research in the later parts and a fuller overview of the area for interested readers.

Within the three parts, each chapter deals with a specific area. To facilitate browsing, a similar layout is adopted in most of the chapters, where possible following the headings: *What is involved? – Why is it important and for whom? – How does it work? – What has been done to date? – What work is needed for the future? – Recommendations.*

1.6 Integrating research in Speech Science and Natural Language Processing (NLP)

The Plan encompasses parallel research in two broad disciplines: (i) Natural Language Processing (NLP), which works with text, i.e. written language; and (ii) Speech Science, which works with sound, i.e. usually the acoustic waveform. While these areas are often pursued separately, close collaboration and integration of these strands will be important, particularly for the applications of Part III.

Natural Language Processing (NLP) entails computational analysis of texts, whether initially composed in written form or as transcripts of spontaneous spoken language. Written language is very prevalent in today's world, and written forms facilitate the analysis of the structural regularities of words, phrases and sentences (generally understood as the *grammar*), and the meaning relationships among them. Many aspects of language analysis are thus carried out using written forms, requiring computing and linguistic skills. These aspects of research are described in Chapter 2.5 and Chapter 2.6. This research also requires large text corpora (Chapter 3), essential for developing the fundamental resources described in Chapters 4, 5 and 6, for text-based language technology building (Chapters 11, 12) and for applications based on them (Part III).

Speech Processing/Speech Science typically entails analysis of the acoustic waveform. The speech signal is highly variable in comparison to written language (Chapter 2.1-2.4). This is due to many factors, e.g. different dialects, the major acoustic differences in the speech of children, men and women etc. Importantly, the spoken message contains dimensions of meaning not represented in text: we interpret meaning, not just from the *words* spoken (which can be written) but also from the *way* they are spoken (which cannot be written). Speech analysis/processing techniques need to handle the acoustic variability of spoken language, capture its multiple layers of meaning, model how humans produce and perceive it, and attempt to replicate this in machines. This requires cross-disciplinary linguistic-phonetic (Chapter 2.4) and speech engineering (Chapter 7) research using extensive speech corpora (Chapter 3) to build speech technologies (e.g. Chapters 8, 9, 10), and applications based on them (Part III).

1.7 Combining knowledge-based and machine-learning approaches

Two complementary approaches, *knowledge-based* and data-driven *machine-learning methods*, feature in the Digital Plan.

Knowledge-based approaches involve the elaboration of explicit knowledge and 'rules' about the language structure, formulated so they may provide the building blocks of a system. Thus, for example, pronunciation dictionaries, letter-to-sound rules, prosody models and morphology rules provide foundation components for technology, e.g. in the building of a text-to-speech-synthesis system (Chapter 8). Part I deals with the provision of such explicit knowledge in technology-available form.

Machine learning entails the use of very large data sets – e.g. matching texts in two languages – and computer algorithms are used to 'learn' the statistical correspondences between them in order to build a machine-translation system. Recent advances in Deep Neural Nets (DNNs) [11] make machine-learning approaches enormously powerful. If vast corpora of the appropriate kind are available, these methods can offer rapid development in certain areas. Machine-learning has to date been crucial in the development of machine translation and speech recognition systems (Chapters 11 and 9) and is increasingly essential to many aspects of technology building.

Both approaches are needed, and each brings specific advantages. A linguistic knowledge-base for Irish that provides a digital, explicit account of the structure of contemporary Irish is an important goal in itself, yielding essential insight into the language. When linguistic resources underpin core technologies they can be continuously improved: as the knowledge base grows and rules are refined, systems can be upgraded. Being technology ready, the knowledge can be used in many applications, and is especially important for educational applications. Intelligent Computer-Assisted Language Learning platforms (Chapter 14) will rely not only on core speech and language technologies, but also on the linguistic knowledge components. Developing a digital knowledge base for Irish is important in many ways for future development but it does demand considerable time and a high level of linguistic expertise.

Machine-learning approaches can offer a quick and less labour-intensive route to developing certain technologies. Deep neural nets are providing impressive results in technology for English. They are also being deployed for Irish technology development and so far results are promising, particularly in light of the development of recent language models for Irish. However Irish, unlike English, does not have the vast corpora and databases or the benefit of many decades of linguistic research – all essential to the effective deployment of machine-learning approaches. In addition, building large models is computationally intensive, requiring significant amounts of computing resources that are not readily available to Irish speech and language technology research. Furthermore, statistically based systems present a 'black box', whose inner workings are not accessible. As new knowledge about the language structure comes to light, these systems cannot be readily adjusted to yield incremental improvements. They also do not yield insight into the language, or linguistic components that can be redeployed, e.g. for educational applications.

In the Digital Plan, these two very different approaches are needed and will often be combined in specific systems.

1.8 A long-term commitment

Although organised as a series of individual chapters/areas, the outputs of the Digital Plan are not envisaged as a set of discrete technology 'deliverables'. Rather, like the tree in Figure 1.1, the various parts are interdependent and should continue to grow and provide new fruit. Specific systems, once built, will need ongoing improvement, with testing, debugging, and evaluation to ensure that they are stable, adapted to user needs and effectively disseminated to the Irish-language community. As the Irish technology sector evolves, new systems and applications will become possible, integrating different dimensions of the research. The Digital Plan entails a long-term commitment, so that as technology evolves, the Irish language is an integral part of it. In line with this goal, the following provisions are recommended, and are discussed more fully in the Conclusions (Chapter 16).

Build infrastructure, nurturing interdisciplinary research centres of excellence, with a high level of Irish: expertise is needed in a number of fields, including linguistics/phonetics, computer science, speech engineering and an advanced level of Irish. In the research carried out to date, recruiting technically skilled researchers with the necessary level of Irish and knowledge of Irish linguistics has proved to be the greatest challenge. There are academic programmes that offer the kind of multidisciplinary education needed, and these programmes should be extended to ensure adequate coverage of Irish linguistic structure and of areas vital to the Digital Plan. Efforts will also be needed to recruit able students who are highly competent in Irish.

Build on existing strengths: although Irish is listed as a language with low provision of speech and language technology [3,4], we are not starting from scratch. As outlined in the various chapters, there are certain areas where there is considerable ongoing activity. The challenge is to extend the scope and scale of research and education to meet the Plan's targets.

Basic and applied research: a balance of basic and applied research is essential (from roots to fruits in Figure 1.1) Basic research provides the foundation for the technologies and ensures their future-proofing, while applied research ensures that research outputs are translated into applications that respond to real-life public needs.

Continuity: more important than the level of funding is the continuity of funding. Without this, research groups will come and go and skilled researchers will be lost as they need to secure long-term employment. Consequently, much of the time and effort invested in research and expertise building is lost. Mechanisms need to be put in place to ensure that research centres working on Irish speech and language technology are maintained and enabled to grow as digital technology grows. Long-term planning is needed to provide a context where young researchers can envisage a future in Irish Research & Development as a viable career option.

Irish-language community - a central partner: the effectiveness of the Digital Plan will depend on the level of partnership with the Irish-language community. Some points are mentioned below in §1.10 and there is a fuller discussion in the Conclusions (Chapter 16) of mechanisms whereby community links might be fostered.

1.9 Global relevance for minority and endangered languages

Building digital capacity promises to be a key factor in the survival of countless languages worldwide: the increasing marginalisation of the languages that do not have digital capacity is accelerating what is already an alarming rate of language loss. There is a growing awareness among the speech and language research community that without positive action it may soon be too late to intervene. This is also reflected by developments such as the *Erasmus+ Digital Language Diversity Project* [12], the *SIGUL Special Interest Group for Under-resourced Languages* [13], the recent EU adoption of the *Language Equality Report* [1] and the *European Language Equality Project* [5].

A guiding principle of the Digital Plan is that the challenges addressed and solutions adopted should be tailored to the Irish context, and for this reason, discussion tends to be localised and detailed. This emphasis on the local context and its specific challenges highlights issues that will resonate with many other endangered language communities. As the types of challenges faced are often very similar, the solutions proposed for Irish are likely to be relevant in many other cases.

As a minority language, Irish is rather unique in having Governmental recognition and support, which prompts the commissioning and implementation of the Digital Plan. In this, we are uniquely placed to assist other linguistic communities. Already, researchers here are networking with other language groups who do not have this level of support. It is a fundamental aspiration of the Digital Plan that our experience, our expertise and, where appropriate, our resources will be shared with other minority/endangered language communities.

1.10 Linkage to the Gaeltacht and Irish-language community

It is of fundamental importance that the Digital Plan should not be divorced from the Irish-speaking community, and particularly from the native-speaker Gaeltacht community. This pertains to every dimension of the research: the building of infrastructure, the training of future researchers, the prioritising of the technologies and applications for development and the dissemination of these technologies as they come on stream. In particular, efforts are needed to ensure that native speakers and competent non-native speakers are well represented among those trained to become key researchers in the requisite technical and linguistic areas.

As the Plan matures, opportunities for downstream commercialisation and campus companies will arise, and forward planning for this will be needed. It is desirable that a policy be developed to regulate how income derived from publicly funded technologies can support the future of the programme.

In the case of campus companies, location in the Gaeltacht (or co-location between Gaeltacht and University/Research Centre) should be encouraged. This will help tap into and support the Gaeltacht community's role in maintaining Irish as a living community language for future generations. This is discussed further in the Conclusions, Chapter 16.

1.11 Our heritage in the digital age

Properly funded, wisely managed, and with an eye to the long-term vision, the Digital Plan has the potential to empower the Irish language community and play an important role in the revitalisation, preservation and documentation of the language. Our language is itself our most precious cultural artefact. It is also key to one of the oldest literary traditions in Europe and to our oral inheritance of song, poetry and story. In bringing the language into the digital age, this Plan aims to seamlessly marry our heritage with modern Ireland, retaining our identity in an increasingly homogenised world culture.

References

[1] Evans, J. *Report on Language Equality in the Digital Age*. retrieved 23 Nov 2019 from http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.html

[2] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy. A. (2015). Speech technology as documentation for endangered language preservation: the case of Irish. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK.

[3] Judge, J., Ní Chasaide, A., Ní Dhubhda, R., Scannell, K. and Uí Dhonnchadha, E. (2012) *The Irish Language in the Digital Age*. Springer.

[4] Lynn, Teresa (2022). *Deliverable D1.20: Report*

on the Irish Language. European Language Equality (ELE). https://www.european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_20__Language_Report_Irish_.pdf

[5] https://www.european-language-equality.eu/

[6] Ó Giollagáin, C. & Charlton, M. (2015). *Nuashonrú ar an Staidéar Cuimsitheach Teangeolaíocht ar Úsáid na Gaeilge sa Ghaeltacht 2006-2011: Príomhfhaisnéis na Limistéar Pleanála Teanga.* Údarás na Gaeltachta.

[7] Government of Ireland. 2010. *20-Year Strategy for the Irish Language 2010-2030.* https://www.gov.ie/en/policy-information/2ea63-20-year-strategy-for-the-irish-language/

[8] Department of Tourism, Culture, Arts, Gaeltacht, Sport & Media. (2018). *Action Plan 2018-2022 for the 20-Year Strategy for the Irish Language 2010-2030.*

[9] Harris, J. (2006). *Irish in Primary Schools: Long-Term National Trends in Achievement.* Dublin: Department of Education and Science.

[10] Houses of the Oireachtas. (2017). *An Caighdeán Oifigiúil: An treoir le haghaidh scríbhneoireacht sa Ghaeilge.* https://data.oireachtas.ie/ie/oireachtas/caighdeanOifigiul/2017/2017-08-03_an-caighdean-oifigiuil-2017_en.pdf

[11] Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning.* Cambridge, MA: MIT Press.

[12] Erasmus+ Programme. *The Digital Language Diversity Project.* https://www.dldp.eu

[13] European Language Resources Association (ELRA), International Speech Communication Association (ISCA). *Special Interest Group: Under-resourced Languages (SIGUL).* https://www.elra.info/en/sig/sigul/

Part I

Resources

Chapter 2
Linguistic Analysis and Digital Documentation of Spoken and Written Irish

2.1 What is linguistic analysis?

Linguistic analysis is the scientific study of how language structure encodes meaning so that it works as a system of communication. An explicit linguistic analysis is needed to develop the most effective technologies to meet the needs of the Irish-language community. For the Digital Plan, the goal is a comprehensive analysis and documentation of Irish structure, using methods that ensure that research outputs are available in forms that can be exploited in speech and language technologies. This linguistic knowledge should also ideally be formulated in ways that can be directly used in applications for the teaching of Irish, for the inclusion of those with disabilities and for the wider public.

This chapter gives an overview of the different elements of language structure to be investigated, indicating what has been done to date. It thus provides a background for the following chapters of Part I, and the broader context for the technology building in Parts II and III.

Spoken and written forms of language employ two different media: sound and text. As mentioned in §1.6, the Digital Plan entails research based both on text (i.e. language composed in written form or transcribed spoken language) using natural language processing (NLP) techniques, and on actual speech data (i.e the speech signal) using speech analysis/processing techniques.

When we talk, the stream of speech is structured in ways, unique to each language, that allow the listener to identify and interpret the meaning of sound sequences, words, phrases and sentences. When we write, we use letters to represent the sounds, spaces to identify word boundaries, commas, full stops and capital letters to identify phrase and sentence boundaries. There are, however, essential differences between written and spoken language, which have implications for how speech and language technologies are developed.

Written language generally provides an idealised version of language, without the hesitations and 'imperfections' that frequently occur in spoken delivery. Text is widely used, particularly in digital communication, and written language is central in education as well as in legal and administrative systems. Text provides a stable and potentially permanent record that can be disseminated, consulted and archived. It greatly facilitates the analysis the language's words, phrases, sentences – aspects often seen as the grammar of the language. In many languages, large repositories of written texts are available and these are important resources for technology building. However, in order to successfully model the structure of a particular language on a computer and fully harness the power of modern computing techniques, one needs not only repositories of data, but also an understanding of the linguistic *structures* (grammar) of the words, phrases, sentences and paragraphs, and how these interact to convey meaning. These structures, which are implicitly understood by humans, need to be made explicit for computational purposes, in the form of annotations[1] and metadata[2]. §2.5 and §2.6 below deal with the prerequisite linguistic analysis of these aspects of language, while Chapters 3, 4 and 5 discuss language repositories (corpora, databases and networks) and the natural language processing (NLP) techniques and tools that are used in language technology development.

Working with the **spoken language** (and the audio signal) is very different. Sounds and words are not discrete and stable as in written form, but vary based on many factors – the context, the speaker, the dialect etc. and speech technology must deal with

[1] Part-of-speech tags, phrasal boundaries, semantic category, etc.
[2] Information about the data such as author, date of publication, speaker, age, dialect, etc.

this inherent variability. Importantly, spoken language conveys many layers of meaning, which are largely unrepresented in written form. Meaning emerges, not just from the *words* and sentences (which can be written), but also from the *way* we say them ***(prosody)*** which is not written. The simple phrase *"I am"* can have many, and potentially opposite, meanings depending on how we say it: confirming, disconfirming, presenting a statement of fact, a question, an order. Our tone of voice establishes the speaker's relationship to the listener (e.g. formal, friendly), conveys the speaker's mood and emotion (e.g. angry, sad) and signals her/his attitude to the broader context of the discourse (e.g. interested, bored). The continuous expressive nuancing that characterises human spoken interaction is of great importance to speech technology. Capturing these wider aspects of spoken Irish for speech technology will require collaborative, interdisciplinary research encompassing linguistic i.e. linguistic phonetics (§2.4) and speech engineering (Chapter 7) using speech corpora (Chapter 3) designed to address specific requirements of the different speech technologies described in Chapters 8-10.

Research on spoken Irish (and Irish speech technology) must also take account of the fact that, unlike the written form, there is no standard spoken form, but rather three main dialects and a number of sub-dialects – all of which are equally regarded as 'gold-standard' spoken varieties. In the major languages, technology has focused on standard dialects, i.e. the dialect of the historically dominant class. For Irish, one must from the outset cater for the diversity of dialects.

Despite inevitable differences in focus, research techniques and the skillsets needed to work on spoken and written forms of language, the two strands are interwoven and outputs are integrated in many of the technologies of Parts II and III.

In the following sections, §2.2 discusses why this work is important while §2.3 gives a brief overview of the main areas of language structure to be investigated. For those interested in further information on the different areas of research and their links to technology, sections §2.4-§2.6 provide more detailed accounts. Further sections deal with language acquisition (§2.7) and language variation (§2.8), both important areas of research for the Digital Plan.

2.2 Why is it needed?

Linguistic research is the foundation for building the ***speech and language resources*** described in Chapters 3-7 that provide the prerequisite components, or building blocks, that underpin the ***core technologies*** of Part II. Knowledge of the structure of Irish is also important to the technology developers, to ensure the technology will be optimised for the language. Developing appropriate speech and language technologies for Irish is not simply a matter of taking off-the-shelf systems, developed for English, and tweaking them at the edges.

The linguistic components are particularly important for many of the ***applications*** envisaged in Part III, through which the benefits to the Irish language community emerge. The range and power of applications is growing by the day, and include applications for the ***public,*** such as comprehensive, well-supported document proofing tools, translation tools and personal assistants with speech and translation facilities (Chapters 11, 12, 13) . ***Educational*** applications are particularly important, as explicit *knowledge* of Irish structure can be integrated into 'intelligent' computer-assisted language learning (iCALL) systems (Chapter 14). The symbiosis of powerful language technologies and digitally available knowledge of Irish structure (§2.4 -§2.6) – along with how it is acquired (§2.7) - have the potential to bring about a paradigm shift in Irish language education. Of fundamental importance are applications for ***disability and access*** to provide adequately for those with speech, language and literacy difficulties (§2.7.3 and Ch.15), and to ensure the full participation of those with disabilities in the world of Irish.

The full digital documentation of Irish language structure is of itself a fundamental goal of the Digital Plan. It will extend our understanding of the language, address aspects not hitherto investigated and embrace the distinct characteristics of the different dialects. It will stimulate research on Irish, and the research outputs will provide unprecedented resources for future scholars, as well as access to state-of-the-art analytic tools.

2.3 How does linguistic structure convey meaning?

A short overview of how different aspects of language structure (areas of linguistics) convey meaning is provided here, with a more detailed account in the following sections, §2.4-§2.6.

Language is a complex system, with many interacting processes occurring simultaneously. For instance, a stream of sound is structured to carry meaning in two ways. At one level, we hear a rapid succession of different ***consonant and vowels sounds***, from which we extract meaningful words, phrases, and sentences. At the same time, the *voice* signal, from which the consonants and vowels are formed, varies continuously with patterns of melody, tone of voice, phrasing, accentuation and rhythm - the ***prosody***. Prosodic structure enables us to decode speech, and group the consonant and vowel sequences into the larger structural units of words, phrases and sentences. Prosody further carries the expressive meanings of speech, the emotion and attitude, not notated in written forms. These aspects (studied under ***Phonetics and Phonology***) are discussed in §2.4.

Individual sounds combine to form ***words***, which are often made up of smaller meaningful parts. For example a word such as *mí-ádhúil* has three parts: *mí-ádh-úil* 'un-luck-y'. Words carry two types of meaning: the ***inherent lexical meaning***, as might feature in a dictionary, and ***grammatical meaning***, telling us how the word relates to other words in the sentence. The morphological makeup of a word may determine its inherent meaning, (*mí-ádh* has the opposite meaning of ádh) or the grammatical meaning (*ádh* 'luck' is a noun, but *ádh-úil* 'lucky' is an adjective). Much of what is traditionally regarded as the *grammar* of Irish is further discussed under ***Morphology*** in §2.5.1 below.

Words combine into larger units commonly known as phrases and sentences. ***Syntax*** (§2.5.2) is the area concerned with the way words and phrases in a sentence contribute to its grammatical meaning. The ***grammatical relationships*** between the words in a phrase (as determined by the makeup of the words themselves and the order in which the words occur), and the relationships between the phrases of a sentence, contribute to the meaning we infer from a sentence.

Syntax interacts closely with ***Semantics*** (§2.6.1) which deals with how the inherent meaning of the words interacts with the grammatical relationships among the words and phrases of sentences. However, meaning emerges not only from the inherent meaning of words and their grammatical relationships, but also from the wider context in which they are used and the way in which they are spoken. This context includes the preceding dialogue, the relationship of the speaker to the listener (or writer and reader), the speaker's (or writer's) intention or state of mind etc. How these all contribute to ***holistic aspects of meaning*** is the area of ***Pragmatics*** (§2.6.2).

The following sections §2.4-2.6 explain these areas in greater detail, pointing out what has been done to date, and relating this research to the development of resources of Part I and the technologies of Parts II and III. For readers who wish to go directly to the technology sections, the extended accounts that follow here will allow for later checking and explanation of specific aspects as they arise.

2.4 Communicating meaning through sound

The sound structuring of Irish is generally little understood, compared to the general understanding of the structure of words/phrases/sentences, understood as 'grammar' (§2.5, §2.6) which is more visible in writing. It is an aspect that is largely neglected in the teaching of Irish and is often the weakest aspect of learners' competence. It is also an area where technology can have a particularly important impact.

Phonetics and **phonology** are the areas of linguistics that deal with the analysis of the speech signal to reveal the patterns that enable the listener to extract the speaker's intended message. Though often viewed as distinct disciplines, the difference is more a matter of perspective. Most researchers working on Irish embrace both aspects, and both perspectives are combined here. In the present document, the terms are used interchangeably, or the term *linguistic-phonetics* is used, embracing both aspects.

2.4.1 Two processes of speech production: two channels of meaning

To understand how speech is structured to carry meaning, one must consider that speech production involves two independent processes, each of which yields distinct patterning in speech, and each of which conveys different aspects of the meaning of an utterance. These two processes are illustrated in Figure 2.1.

(a) Sound Generation:
Voice results when the vocal cords vibrate, converting air from the lungs into an acoustic signal, the carrier wave of speech. As we speak, the voice signal varies constantly, imposing patterns of melody, tone-of-voice, accentuation, phrasing and rhythm (these patterns are called prosody).

(b) Sound Filtering:
The voice signal is filtered by the vocal tract as it travels from the vocal chords to the lips. The

continuous movements of the tongue and lips (articulation) change the shape of the filter, thereby altering certain acoustic properties of the voice signal, differentiating it into the individual consonant and vowel sounds of the language. The underlying prosodic patterns remain intact.

Thus, the voice is the sound source for vowels and consonants, and they are differentiated by different filtering effects of the vocal tract[3]. The listener accesses the meaning of the words from the rapid sequencing of of vowels and consonants (§2.4.2) while the prosodic patterns give access to the wider meanings of the sentence (§2.4.3).

Traditionally, trained phoneticians provided auditory, impressionistic transcriptions of the sounds and prosody of the language, using internationally agreed conventions [1] to describe a given language: in [2,3] these conventions are used to describe the consonants and vowels of Irish. Modern digital techniques add greatly to what the ear can analyse: they provide acoustic characterisation of the sounds and prosody, and direct analysis of articulation (movement of speech organs) can further illuminate how the sound system works. When working directly with speech signals, it is clear that the sounds (consonants and vowels) are not discrete (as written letters imply) but overlap considerably so that acoustic characteristics (and articulatory gestures) for one sound spill into the neighbouring sound. Sound is therefore very variable, being heavily influenced by preceding and following sounds and by the prosodic context. The acoustic frequencies of a given sound are also dependent on the size of the speaker's vocal folds and vocal tract, and are therefore very different for children of different ages, women and men. These sources of variability present no difficulty for the human listener, but they do for machines, and are important considerations in the design of synthesis (Chapter 8) and recognition (Chapter 9) systems. For Irish, dialect variation is a central concern for the linguist and for the development of speech technology.

To be useful for the Digital Plan, the linguistic-phonetic analyses must be carried using methodologies that provide the kinds of research outputs that can be exploited in technology. Close collaboration with engineering modelling approaches (Chapter 7) is important, to allow implementation in technology and direct transfer of knowledge in applications, e.g. for language teaching.

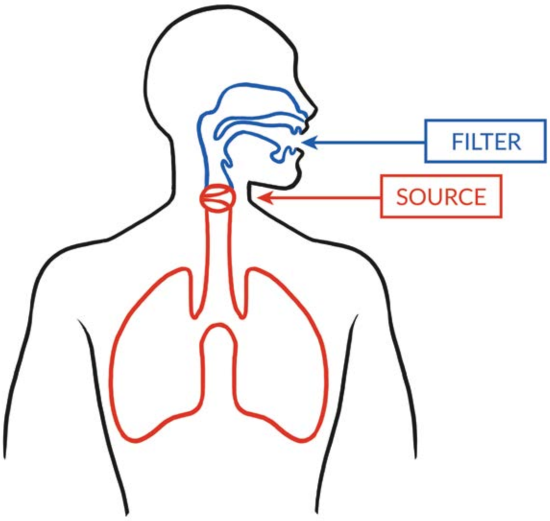


Figure 2.1: Two process of speech production:
(a) sound generation (voice source) and
(b) sound filtering (articulation)

2.4.2 Articulation of Irish Consonants and Vowels (Sound Filtering)

Over the years, the main focus of phonetic-linguistic research on Irish has been the articulation of consonants and vowels, and there is a long, rich tradition of impressionistic, auditory analysis of different dialects, e.g. [4-12]. The consonantal system is extensive, with a fundamental distinction between two consonant qualities: the broad and slender, for example the different initial consonants /lˠ/ agus /lʲ/ in the words /lˠo:nˠ/ *lón* 'lunch' and /lʲo:nˠ/ *leon* 'lion' [2]. Neither of these two are the same as the English /l/, as in /lo:n/ 'lone'. The broad and slender consonants exert a strong influence on the neighbouring vowel. The large consonantal system has implications for the design and size of speech corpora that are needed for technology: given how strongly a sound influences its neighbours, one needs recordings of every sound in every possible context it occurs in.

The broad-slender contrast can differentiate words' meanings, as in the above example. It can also affect the grammatical meaning of the word: for example, when the final consonant in the word *leon* changes from /nˠ/ to /nʲ/ (/lʲo:nˠ/ *leon* to /lʲo:nʲ/ *leoin*) the word can change from singular to plural or from nominative to genitive (see [2] and §2.5.1 below).

There has been relatively little empirical analysis with modern analytic techniques of the articulatory and

[3] This account is simplified to highlight the basics.

acoustic features of Irish consonants and vowels. Such data is important, especially given how different the system is from English, for which most current technologies have been designed. A number of studies with direct recordings of articulation, and some acoustic analyses [13-19] help illuminate the broad-slender system of contrasts as well as some cross-dialect differences [20-22].

Looking to the future, for the collection and analysis of articulatory/acoustic data that can effectively be deployed in speech technology, extensive collection and processing of data using state-of-the-art imaging techniques like fMRI would be optimal. This can provide a deeper insight into the language, illuminate cross-dialect differences and provide a basis for core technologies and for applications that hold great promise for Irish-language teaching and therapy. As mentioned above, such methodologies will require a linguistic-engineering collaboration (see elaboration in Chapter 7).

2.4.3 Prosody of Irish: The Voice Source

Prosody is about how we vary the sound of the voice to form meaningful patterns – in melody, tone-of-voice, in the accentuation and rhythmic sequences of syllables. Prosody provides different kinds of information in spoken communication. Firstly, it enables the listener to ***decode*** the stream of speech – finding the words, phrases etc. It allows us to identify which words are most important in a sentence as well as other information on how it is to be interpreted. Much of this information is not notated in written forms. Secondly, as discussed above, tone-of-voice conveys many other layers of meaning, expressing the speaker's ***feelings*** and ***attitudes*** – aspects not indicated in writing, but of great importance to how humans communicate. Prosody also serves to regulate conversation, as the speaker signals whether she/he is yielding or holding the floor, also an aspect not notated in writing.

Until relatively recently, there was little coverage of Irish prosody, compared to the work on the vowel and consonantal sounds. Impressionistic analyses of Connemara melodic patterns (intonation) were provided [23, 24], and there are some isolated observations in earlier dialect descriptions [6, 7]. More recent, empirical acoustic studies have been shedding light on the melodic and accentuation patterns of the Irish dialects. There are striking cross-dialect differences, most notably a major North-South divide, with predominantly rising tones in Donegal, but predominantly falling tones in Connaught and Kerry [25-30]. Even closely related dialects, such as Inis Oírr and Cois Fharraige can exhibit striking differences in prosody [31]. This work highlights the need for multi-dialect coverage.

The studies to date have mainly focused on the melodic and accentuation patterns and their *encoding* functions. For the future, such work needs to be extended to include analysis of *tone-of-voice* in expressing emotions and in regulating human conversational interaction. This will be critical for future speech technologies, where we will be conversing with virtual characters, robots etc. and where an approximation of human prosody is essential (see Chapters 8-10). To provide a fuller account of Irish prosody, joint research with speech engineers is needed, using state-of-the-art modelling techniques (see Chapter 7). As in the case of articulatory analyses, such an interdisciplinary approach is optimal, both to facilitate analyses that extend our linguistic understanding of Irish prosody, and to ensure this knowledge can be translated to the technology and embedded in applications for the Irish language community.

2.4.4 Mapping sounds to written letters

In alphabetic writing, the letters represent the consonant and vowel sounds, while prosody is largely not notated. In Irish, the mapping of consonant and vowel sounds to the corresponding letters is complicated and opaque. It is complicated in that a given sound can be written down in many different ways. The writing system is opaque in that the distinct broad-slender consonantal sounds are not immediately obvious. (Note however, though complicated, the mapping between sounds and letters is much more regular in Irish than in English). These features of the writing system have implications for the acquisition of literacy, and these are discussed further in §2.7.3.

Current speech synthesis and recognition systems typically use text as input/output (Chapters 8, 9), and therefore, the letter-to-sound correspondences need to be elaborated. There are considerable cross-dialect differences, and this mapping must be done for each dialect. Letter-to-sound rules have been developed for the three main dialects of Irish, as part as part of the development of synthetic voices in the ABAIR[1] initiative at Trinity College, Dublin. This work is being exploited also for research on literacy acquisition [32] (see §2.7.3 and Chapter 15).

[1] ABAIR is supported by the Department of Tourism, Arts, Culture, Gaeltacht, Sport and Media.

2.5 Communicating meaning through grammatical structure

Morphology and **syntax** are the areas of linguistic analysis which concern the structure and meaning of words, phrases and sentences – areas which are included in traditional grammatical accounts, and where analysis is typically based on written forms.

2.5.1 The structure of words: morphology

In terms of language technology, it is fundamentally important to be able to recognise the words (and potential words) of the language. It is also necessary to be able to identify the features of a word, such as its root form (headword or lemma), and what type of word it is, i.e. whether it is a verb or a noun etc. (part-of-speech category). In the case of verbs it is important know the tense (past, present, future etc.), and other features such as number, person, aspect, mood etc. are also necessary. In the case of nouns, it is important to know whether the noun is singular or plural, and other features such as gender (feminine or masculine) and grammatical case (nominative, genitive, vocative etc.) are also important. This kind of information is essential for many common language applications, such as spellcheckers, grammar-checkers, dictionaries, thesauri (Chapter 13), as well as for search engines (Chapter 12), translation applications (Chapter 11), educational applications (Chapter 14) and language generation (Chapter 6).

Linguistic morphology is the study of the internal structure of words. Each language has its own system of rules for introducing new words to the language, and for creating the various forms of words. Usually the system involves adding affixes (e.g. suffixes and prefixes) to the root or altering the root or both. Therefore, many words are composed of smaller identifiable subparts, i.e. the root and prefixes and/or suffixes. These sub-parts are known as 'morphemes'. For example, the word *tógfaidh* 'will take' has two morphemes: the lexical root *tóg* and the future tense suffix *-faidh*. The root itself can also change. Root changes include initial mutations such as *séimhiú* (lenition) e.g. to mark past tense; *thóg* 'took' vs. *tóg* 'take', and internal modifications such as *caolú* (slenderisation – the switch from broad to slender final consonant, e.g. to mark the plural form; *báid* 'boats vs. *bád* 'boat' (see §2.4.2).

Because the basic root of a word (lemma) routinely changes shape, and especially because the start of the word can change, natural language processing (NLP) tools such as lemmatisers which identify the root (i.e. the lemma), and part-of-speech (POS) taggers which identifty the (POS) category and associated features, are particularly important for Irish speech and language processing.

The phenomena of Irish morphology are well documented [33, 34] and rule-based morphology tools and POS taggers have been developed (§5.4) which can identify/generate the lemma and features of a word and assign POS tags to a sentence or text [35]. The main challenge for Irish morphological analysis in the context of language technology is keeping the system constantly updated as new words come into the language. This can be achieved by establishing strong links with lexical databases (Chapter 4). The rule-based morphology tools could be enhanced through incorporating more multi-word items, named entities such as names of people, places and organisations (§5.5) and common word co-occurrences (§3.2).

In situations where language does not always conform to the standard morphology rules, e.g. social media data, statistical (rather than rule-based), corpus-based methods can be used. In terms of statistical language processing and machine learning, the rich morphological system of Irish, where a single word may have many spoken and written forms depending on the grammatical context, presents a challenge. The term 'data sparsity' is used to describe the link between morphology and the impact it has on required corpora sizes in statistically based methods (see §2.7). This means that in building corpora (Chapter 3) for statistical or machine learning, e.g. statistical machine translation (Chapter 11) or speech recognition (Chapter 9), corpora need to be very large in order to provide sufficient coverage of all of the word-forms of the language.

2.5.2 The structure of sentences: syntax

Quite often, it is not enough to only know about the properties of the words in a sentence, we also need to know what their role is in a sentence. This is important for the basic understanding of 'who did what to whom'. Syntax is concerned with sentence structure, and, in particular, the system of rules that determine how words are combined to form phrases and and how phrases are combined to form sentences. It is also concerned with identifying grammatical functions (§3.4.1 written corpora) such as subject and object, and in determining the hierarchical relationships between phrases in a sentence. This type of analysis is known as parsing, and a computer program that performs automatic syntactic analysis of language is known as a parser. Parsing is of central importance in many natural language processing applications, such as grammar checking, intelligent CALL applications, rule-based machine translation, and natural language understanding and generation. Parsing builds upon morphological analysis and part-of-speech tagging as mentioned in the previous section.

(1) Léigh an cailín an litir
Read the girl the letter
'The girl read the letter'

We can often identify the grammatical subject and object of a sentence in Irish (and English) by their position in the sentence. For example, sentence (1) consists of the verb, *léigh* 'read' followed by two noun phrases; *an cailín* 'the girl, and *an litir* 'the letter'. In Irish, we know that the first noun phrase following this verb is the grammatical subject and the second noun phrase is the grammatical object.

(2) Léigh an cailín clúdach na litr-each
Read the girl cover the letter-GENITIVE
'The girl read the envelope' i.e. the cover of the letter

In sentence (2) we know that *clúdach na litreach* is a single entity, i.e. a single noun phrase with the grammatical function 'object', because of the genitive case (*an tuiseal ginideach*) relationship between the nouns *clúdach* and *litreach*.

In NLP, being able to identify syntactic relationships is important for interpreting the meaning of a sentence. It is essential for information retrieval and document summarisation (Chapter 12), and it is particularly important for language pairs such as English and Irish, which have different word order. Parsing and chunking tools for Irish are described in Chapter 5.

In terms of Irish syntax, there is surprisingly little detail in traditional grammar reference works [33, 36]. The majority of theoretical studies of Irish syntax have been carried out within the Chomskian generative framework [37-41], and Role and Reference Grammar (RRG) functional framework [42]. These frameworks were not designed for computational purposes, therefore a number of computationally oriented theoretical models of grammar and syntax have been developed over the last few decades, notably Lexical Functional Grammar (LFG) [43] and Head-driven Phrase Structure Grammar (HPSG) [44]. These syntactic frameworks recognise the need to integrate various levels of linguistic analysis, including morphology, semantics and pragmatics, and present their analyses in a computational manner. Some aspects of Irish syntax have been addressed using the LFG framework [45, 46]. In recent years, dependency syntax [47] has also grown in popularity as a computational framework of syntax.

While much research has been carried out on the theoretical syntax of English and other major European languages, this does not necessarily carry over to Irish syntax because of the significant differences between the languages in terms of word order and idiom. Further research on the syntax of Irish, particularly within a computational framework is essential, in order to provide a solid basis for robust computational parsing, for reference grammars for the language and for CALL applications etc.

2.6 Inherent and holistic meaning

The study of semantics is concerned with the inherent meaning of words and their relationships to one another (i.e. the information stored in lexical knowledge bases as detailed in Chapter 4), while pragmatics is concerned with the ways in which meaning is affected by context. In practice we have a more holistic attitude to meaning, in that we interpret words and sentences differently in different contexts. In spoken conversation, the prosody, and particularly the tone-of-voice used by the speakers provides a great deal of the overall meaning.

2.6.1 Inherent meaning: semantics

In order for a computational system to interpret a sentence, i.e. assign a particular structure to a sentence, it is necessary to have semantic information relating to the meaning of words. For instance for a particular verb it is useful to know how many participants are typically involved in the action, i.e. the verb's predicate-argument structure. This will help determine how many noun phrases we can expect in the sentence. It is also useful also to know what roles these participants fulfil, as this will map on to the grammatical functions of subject and object etc. Information about verbs in terms of the number of participants they require, and the semantic roles required is encoded in a valency dictionary. Information about what kinds of nouns that can fulfil those semantic roles is known as the predicates' selectional restrictions or preferences. For instance a verb such as 'drink' needs a direct object, but we can also say that that object should be a liquid. In addition to the number of entities/participants required, the interpretation of a sentence can depend on the characteristics of the entities themselves.

(3) Tá sí ina gairdín
Is she in.her garden
'She is in her garden'

(4) Tá sí ina dochtúir
Is she in.her doctor
'She is a doctor'

For instance, sentences (3) and (4) have a similar sentence structure but we interpret them differently based on the characteristics of the nouns involved. *Gairdín* 'garden' is interpreted as a location while *dochtúir* 'doctor' is interpreted as a profession, i.e. the nouns have different semantic properties. Therefore, it is also necessary to document information about categories of nouns. Knowing whether a noun is human, animate, inanimate, liquid, solid, a location etc. facilitates robust syntactic parsing and semantic interpretation of meaning. In addition, systematic semantic relationships such as synonyms (words with similar meaning), antonyms (words with opposite meaning) and hypernyms/homonyms (classes and subclasses) can be encoded in knowledge bases (also known as 'wordnets'), in order to provide valuable semantic information for NLP applications (see Chapter 4.3).

Very little linguistic research or documentation has been undertaken to date on the semantics of Irish. The type of semantic research that is required for speech and language technology needs to take place in the context of suitable computational frameworks. It needs to be corpus based (Chapter 3) and it will require specialist language processing tools (Chapter 5). In subsequent chapters, we describe semantically annotated corpora (Chapter 3), valency dictionaries and semantic networks (Chapter 4) and annotation tools such as semantic role labellers and named entity recognisers (Chapter 5).

2.6.2 Holistic meaning: pragmatics

The meaning of a sentence or utterance is not just the sum of the individual words, the syntactic structure or the inherent semantic meaning. Meaning also depends on the context of use, and in the case of spoken language, the speaker's voice prosody. Pragmatics looks at the more holistic aspects of meaning, and operates at the level of 'discourse', where sentences or spoken utterances combine to form a coherent text or dialogue. The relationship between the speaker and hearer, or the writer and reader are important considerations in the study of pragmatics, as this will influence the degree of formality that is appropriate. The meaning in a particular context will also depend on the participants' prior knowledge and assumptions, as well as general 'world knowledge'. Much international research has centred on finding appropriate and computationally efficient ways of encoding world knowledge to assist in natural language understanding, however very little research on the pragmatics of Irish has been carried out to date.

In written and spoken discourse, repetition is generally avoided for economy and to highlight the difference between new and previously mentioned information. This often means using pronouns in place of recently mentioned nouns. One of the challenges in natural language understanding applications is to correctly interpret the pronouns, i.e. determining which noun phrase (or statement) a pronoun refers to. The answer is usually to be found in the text (discourse) or in the domain of general knowledge. For instance, in *Chonaic mé é*, 'I saw it/him' the pronoun *é* refers to some previously mentioned person or thing, i.e. this sentence could be interpreted as 'I saw it' or 'I saw him'. The discourse which precedes this sentence, together with world knowledge and our shared assumptions, will determine the most likely referent for the pronoun in this particular context. Similarly, in a sentence such as 'The president made a speech', 'the president' could refer to the president of a country, organisation or a company, and it is also time-dependent. Determining the referents of pronouns in a text or transcript, is known as 'anaphora resolution'. Knowledge of the principles of discourse structure and pragmatics for Irish is a prerequisite to developing the kind of tools which can perform this NLP task (Chapter 5). Anaphora resolution is important for Natural Language Understanding, particularly in Educational applications (Chapter 14) and for Information Retrieval (Chapter 12)

In spoken interaction, voice prosody (§2.4.3 above and Chapter 7) is a primary carrier of pragmatic information. For example, prosodic focus conveys which part of an utterance is important and which part is not (i.e. prior knowledge). How an utterance is to be interpreted (e.g. as a statement, question, order or exclamation) is also transmitted through the voice. Prosodic signals regulate the conversational flow, signalling where the listener takes a conversational turn. And most fundamentally perhaps, the prosody (especially tone-of-voice) marks the speaker's relationship to the listener (e.g. *formal*, *intimate*, or *condescending*). In conversation speakers' prosody converges and diverges, reflecting the evolving connection between them and signalling their feelings (*sad*, *excited*, *bored*). A 'mismatch' of words and tone-of-voice can signal sarcasm. These aspects of meaning are essential for effective spoken communication, and for enriched speech technology applications (see Chapters, 7, 8, 10).

In the area of language teaching and assessment (§2.7.2), high pragmatic competence is an indication of high language proficiency. Therefore being able to recognise pragmatic constructs, e.g. appropriate use of idioms, indirect requests, and other appropriate discourse features is important for language teaching and assessment. Annotated written and spoken corpora (§3.3) can be analysed in order to identify pragmatic constructs. In addition, annotated learner corpora (§3.5) are indispensable for tracking the development of pragmatic competence in learners over time. Learner's acquisition of native-like prosody is essential to their pragmatic competence in spoken language. Coordinated cross-disciplinary research is required to ensure that phonetic knowledge (§2.4.3), can be translated into teaching-friendly form, and embedded in pedagogical applications (Chapter 7, 14).

Speech and language technology applications that need to generate language, either as text or speech, must produce language which is both grammatically accurate and pragmatically appropriate. Correct interpretation and generation of the discourse is very important for spoken and written dialogue systems (Chapter 10), machine translation (Chapter 11) and question answering systems (Chapter 12). For interactive spoken dialogue systems, which are increasingly approximating human spoken interaction, voice prosody (§2.4.3) is particularly critical (Chapters 7, 8, 10). Pragmatics is therefore an essential part of the more sophisticated speech and language technology applications.

To date there has been very little research regarding the pragmatic aspects of Irish language use, either in written or spoken language. Text-based pragmatic studies of communication need to focus on the pragmatic elements of written discourse and transcribed spoken discourse. Complementing such text-based analyses, for speech technology to access the holistic meaning in spoken interaction, research is needed on the pragmatic dimensions carried by voice-prosody. Fundamental research is needed, as pragmatic awareness is important to a broad range of speech and language technologies. Research on pragmatics can also support curriculum design, teaching and assessment for Irish.

2.7 Language and Literacy Acquisition and Proficiency levels

Research into Irish language acquisition is an essential basis for linguistically informed educational technologies. In addition to language technology, such research will have many other important benefits for Irish language teaching and for the diagnosis of language impairments in the context of acquiring Irish as a first language.

The process by which a first language is learned in early childhood, and the way in which a second language is learned in an educational setting, are different in several ways. This has led to separate fields of study for first language acquisition, (with sub-fields for monolingual, bilingual and multilingual acquisition), and for second language acquisition.

2.7.1 First Language Acquisition & Bilingual First Language Acquisition

One of the ways in which first language acquisition is studied is through the collection of samples of spontaneous spoken language for analysis. Typically, recordings are made of interactions between a child and their carer at regular intervals in a natural setting during the first few years of a child's life. A number of studies of childhood first language acquisition of Irish have taken place over the years [48-51] and valuable data has been collected. Nevertheless, compared to other languages such as English, relatively little is known about when and in what order the different aspects of the sound system, the morphology and grammar of Irish are acquired by the native speaker. Too often assumptions are made that the process of acquisition of Irish (as a first or second language) will mirror that of English or other well-studied languages. This assumption is not at all warranted given the many large structural differences between languages.

Targeted empirical analyses of first language acquisition data are required in order to chart the typical progress of acquisition of Irish as a first language in a bilingual/ monolingual environment and the relevant milestones. This will involve the collection of addition data, to complement existing data. This basic research is an essential foundation for effective CALL systems and such research can also underpin language curriculum design for Irish medium schools, in both Gaeltacht schools and Gaelscoileanna.

2.7.2 Second Language Acquisition and Language Proficiency Levels

A learner's experience of learning a language is inevitably influenced by their first language. The degree of difficulty the learner experiences in acquiring a second language, is partly related to how different the second language is from their first language. Usually, the features of a second language not present in the first language will present difficulties for a learner. Irish and English, as already alluded to, are significantly different. One of the ways of meeting the challenges that this presents is through the detailed analysis of second language learners' progress, through the collection and analysis of learner corpora (§3.5) at the various levels of education.

As the majority of people acquire Irish as a second language, typically after English has been acquired monolingually, there is a pressing need for research in this area to inform the development of appropriate educational technologies. In addition to strongly grounded educational technology systems, this research will also support curriculum design for teaching and assessment of Irish in mainstream schools.

Research into the acquisition of Irish as a second (or subsequent) language in a school setting or among adult learners is surprisingly scant. One of the ways in which second language acquisition is studied is through the analysis of learner language corpora [52] (§3.5), and further learner corpora in a variety of settings will be needed. Much research on second language acquisition is currently taking place in the context of the Common European Framework of Reference for languages (CEFR) [53-55]. CEFR uses the levels A1 and A2 (basic users), B1 and B2 (independent users) and C1 and C2 (proficient users) and it recognises the separate and distinct skills of speaking, listening, interacting, reading and writing. It is common for language learners at a particular point in time to have different levels of proficiency in these different skills.

The Common European Framework of Reference (CEFR) proficiency levels A1 to C2 are specified in the framework in a language independent manner using 'Can Do' statements . This means for a specific language, such as Irish or English etc. the 'Can Do' statements for each CEFR level must be mapped to specific elements of phonology, morphology, syntax, lexicon and discourse etc. in the language. The mapping of Can Do statements to specific elements of the language can be carried out most efficiently using digital resources and tools. Detailed corpus analyses can be used to benchmark the CEFR levels for Irish, by determining the specific vocabulary items, grammatical structures and spoken language features, which can be associated with each level of proficiency.

Innovative corpus-based research into CEFR levels and Irish is being carried out in Trinity College, Dublin and in Maynooth University, using corpora of teaching materials [56], and written and spoken learner corpora [52], together with general reference corpora (Chapter 3).

As well as the importance of CEFR benchmarking for the development of pedagogically sound CALL applications, this research is also vitally important for the development of integrated language curricula and assessment which will support and enhance the teaching and learning of Irish at all levels of education. Chapter 3 details the types of learner and proficient-speaker corpora that are required for research into first and second language acquisition for Irish, together with the language annotation and analysis tools (Chapter 5) that will be required to support this research.

2.7.3 Literacy Acquisition

While we learn to speak quite naturally, learning to read and write are skills that need to be taught. Essential prerequisites of literacy acquisition are a grasp of (i) the sound contrasts of the language (an essential part of phonological awareness) and of (ii) how the sounds map to written letters (phonic awareness). The transparency of the writing system, i.e. the sound–to-letter mapping, varies from language to language. When simple and consistent, as in Spanish, children learn to read much more quickly than where the mapping is complex and/or irregular, as in English [57].

Irish has (i) a rich sound system, with a contrast of broad/slender consonants that does not exist in English (§2.4.2 and [2]), and (ii) a complex mapping of the sounds to letters (phonics system), which is also very different from English, though much more regular. The complex sound-to-letter mapping of Irish stems partly from the fact that the (Roman) alphabet does not provide a way of marking the broad/ slender consonantal contrast on the letters indicating consonants. The consonant quality is indicated rather through the choice of neighbouring vowel letters, so in the word 'buíon' *band*, (pronounced /bˠi:nˠ/), the <u> and <o> letters are used to signal the broad initial and final consonants [2]. The opaqueness of the writing system is also partly due to its antiquity: pronunciation changes over time, spellings don't tend to keep up with shifts in pronunciation. Futhermore, pronunciation changes evolve differently in individual dialects: consider modern pronunciations of *(ní) bhfaighfidh* '(not) get' (future tense) → /ʋˠaɪgʲ/ (Kerry), /wˠaɪ/ (Conemara), /wˠi:/ (Donegal).

Nonetheless, the fact that the mapping of sound to letter in Irish is fairly regular means that a phonics-based approach to literacy training should be optimal. In spite of this, whole-word memorisation appears to be the mainstream approach [58, 59]. This is far from ideal, especially as word forms in Irish tend to vary a great deal (see §2.5.1).

For learners who have already learned (or are simultaneously learning) to read and write English, the large differences in the sound system and the spelling rules of the two languages must be borne in mind. Use of English alphabet pronunciation, and a frequent assumption that sounds can be 'transferred' from English to Irish, are unhelpful, obscuring the native contrasts of Irish and more likely to undermine Irish literacy acquisition. Clearly, the two languages should ideally be taught in ways that respect the phonic and and writing systems of each one.

Despite a growing awareness of the need for a phonologically informed approach [59], there are few resources available to learners and their teachers to develop phonological awareness, the bedrock for developing phonics awareness. With some notable exceptions, such as the excellent work of [60, 61], most current teaching materials are modelled on English and are therefore far from ideal [62]. In the case of learners with dyslexia, it is particularly unsatisfactory that there are no diagnostic materials or remediation tools available. Those designed for English are entirely unsuitable and learning support teachers and educational psychologists are left to manage as best they can [63, 64].

The need for Irish resources has been acknowledged in the *Polasaí don Oideachas Gaeltachta* 'Policy for Gaeltacht Education' [65]. These must be designed from the ground up and founded on a full understanding of the phonology and phonics of the language, and cannot be simply a remodelling of resources designed for English. Towards such goals, basic research on the acquisition of phonological awareness and phonic principles in Irish has been investigated for both native- and second-language learners at Trinity College, Dublin [66-68]. Linked to this research, the ABAIR - CabairE collaboration in TCD is developing digital interactive games to train phonological awareness and early literacy skills ([69] and Chapter 14). While aimed at all learners, these are particularly needed for learners with dyslexia (Chapter 15). Looking to the future, online literacy screening should be possible, as well as online training and remedial resources, drawing on the linguistic knowledge and exploiting the interactive multimodal technologies of the Digital Plan.

To sum up, digital technology can make a positive contribution to the acquisition, teaching and assessment of Irish by (1) using digital language resources and tools to define the milestones of first language acquisition and second language learning, and (2) developing computer-assisted applications, informed by these milestones to support all learners, including those with reading difficulties and with speech and language difficulties.

2.8 Language Variation and Language Change

Like all living systems, language shows considerable variation across dialects, social groups, ages of speakers and contexts of use. Language is also inevitably subject to change over time and contemporary dialect variation is often the reflection of different changes taking place at different rates across geographical locations. Contact with another language is often a source of language change. Language contact naturally has implications for first and second language acquisition of Irish, given that Irish is a minority language existing in close proximity with a global language like English. Language variation and language change can be researched using contemporary and historical corpora of Irish (§3.6), both written and spoken.

Written language tends to vary less than the spoken language, given that there is an established written standard, and given the fact that written language tends to be conservative vis-à-vis changing spoken norms. Nonetheless, in an increasingly online world, it is important to include study of more variable written forms also, such as the language used in social media in the daily lives of Irish language users, in addition to the edited text found in literature or news.

Research on language variation and language change should be taken into consideration in language standardisation policies, language planning, the creation of reference grammars and writing standards. These language resources all feed into digital technologies such as spellcheckers, grammar-checkers (Chapter 13) and educational applications (Chapter 14).

As discussed in §2.4 above, language variation is a central preoccupation in designing speech technology. Research on cross-dialect variation is essential for dialect-appropriate technologies, including speech synthesis (Chapter 8), speech recognition (Chapter 9) and spoken dialogue systems (Chapter 10), as well as applications that are based on them. Other types of variation, such as age and gender, are also essential considerations in the design of speech-based technologies and applications.

2.9 Recommendations

- Undertake extensive empirical, linguistic and computational research in the various aspects of Irish language structure, based on both spoken and written resources. This work should prioritise (i) aspects where there are gaps in our current knowledge of language structure, with implications for technology development, and (ii) acquisition of Irish as a first and as a second language.
- Ensure that the research methodologies used:

a. enable the resource developments in Part I
b. can be implemented in core technologies in Part II
c. can be directly exploited in applications (Part III) that support Irish language learning; support those with difficulties in speech, language or reading; and support the wider Irish language community.

- Foster collaborative interdisciplinary research to facilitate 1 and 2. Ideally, linguists will work closely with engineers and computer scientists, providing linguistic resources for technology and exploring together how knowledge of the language can be maximally exploited in applications.
- Train appropriate researchers: this requires undergraduate and postgraduate programmes for the training of researchers with the requisite (Irish) language, linguistic and technical skills. Programmes should provide training in digital empirical linguistic research methods, and an understanding of the speech and language technologies they are aimed at. This could be done by expanding/adjusting existing programmes to help them more fully meet the requirements of the Digital Plan, and/or developing new programmes as appropriate.
- Provide continuous research support: this requires recurrent research funding for postgraduate and (where possible) postdoctoral researchers to work in the requisite key areas, to build and maintain a lively Irish speech and language technology sector.
- Provide continuous support for the provision of the requisite corpora for linguistic research and technology development. This will include (i) updating and making existing corpora more widely available and user-friendly; (ii) development of additional corpora (see Chapter 3).

References

[1] *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet.* (1999). Cambridge: Cambridge University Press.

[2] Ní Chasaide, A. (1999). Irish. In *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet.* Cambridge University Press, pp.111-116.

[3] Ó Raghallaigh, B. (2014). *Fuaimeanna na Gaeilge.* Dublin: Cois Life.

[4] Quiggin, E. C. (1906). *A Dialect of Donegal.* Cambridge University Press.

[5] Sommerfelt, A. (1922). *The Dialect of Torr, Co Donegal*. Christania: Jacob Dybwad.

[6] Ó Cuív, B. (1944). *The Irish of West Muskerry, Co. Cork: a phonetic study*.

[7] de Bhaldraithe, T. (1945). *The Irish of Cois Fhairrge, Co. Galway: a phonetic study*. Dublin Institute of Advanced Studies.

[8] Breatnach, R.B. (1947). *The Irish of Ring, Co. Waterford: a phonetic study*. Dublin Institute of Advanced Studies.

[9] de Búrca, S. (1958). *The Irish of Tourmakeady, Co. Mayo: a phonemic study*. Dublin Institute of Advanced Studies.

[10] Mhac an Fhailigh, É. (1968). *The Irish of Erris, Co. Mayo: a phonemic study*. Dublin Institute of Advanced Studies.

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

[12] Ó Sé, D. (2000). *Gaeilge Choirce Dhuibhne*. Dublin Institute of Advanced Studies.

[13] Bennett, R., McGuire, G., Ní Chiosáin, M. & Padgett, J. (2017). An ultrasound study of Connemara Irish palatalization and velarization. *Journal of the International Phonetic Association 48*(3), 261-304.

[14] Padgett, J. & Ní Chiosáin, M. (2018). The perception of a secondary palatalization contrast: A preliminary comparison of Russian and Irish. In R. Bennett, A. Angeles, A Brasoveanu, D. Buckley, N. Kalivoda, S. Kawahara, G. McGuire & J. Padgett (Eds), *Hana-bana (花々): A Festschrift for Junko Ito and Armin Mester.* Linguistics Research Center, University of California Santa Cruz, pp. 158-171.

[15] Ní Chiosáin, M. & J. Padgett, 2012. An acoustic and perceptual study of Connemara Irish palatalization. *Journal of the International Phonetic Association 42*(2), 171-191.

[16] Fitzpatrick, L. (2000). *Analysis of Irish Lingual Articulation Using Electropalatography and Electromagnetic Articulography*. [Ph.D. thesis]. Trinity College, Dublin.

[17] Ní Chasaide, A. & Fitzpatrick, L. (1995). Assimilation of Irish velarised and palatalised stops. In *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. 334-337).

[18] Ní Chasaide, A. & Fealy, G. (1991). Articulatory and Acoustic Measurements of Coarticulation in Irish (Gaelic) Stops. *In Proceedings of the 12th International Congress of Phonetic Sciences.*

[19] Ní Chasaide, A. (1977). *Acoustic study of the laterals in Donegal Irish and Hiberno-English*. [M.A. thesis]. University of North Wales, Bangor.

[20] Ní Sheighin, M. (2001). *An Electropalatographic Study of the Consonants of the Dialect of Irish of Dú Chaocháin, Co Mayo*. [M.Phil. thesis]. Trinity College, Dublin.

[21] Russell, É. (2001). *An EPG Study of the Irish Dialect of Coirce Dhuibhne*. [M.Phil. thesis]. Trinity College, Dublin.

[22] Cooper, A., (1994). *An EPG Study of the Irish Consonants of Cois Fharraige*. [M.Phil. thesis]. Trinity College, Dublin.

[23] Blankenhorn, V. (1981). Intonation in Connemara Irish: a preliminary study of kinetic glides. *Studia Celtica 16-17*, 259-79.

[24] Bondaruk, A. (2004). The inventory of nuclear tones in Connemara Irish. *Journal of Celtic Linguistics 8*, 15–47.

[25] Dalton, M. (2008). *The phonetics and phonology of the intonation of Irish dialects.* [Ph.D. thesis]. Trinity College, Dublin.

[26] Dalton, M. & Ní Chasaide, A. (2007). Nuclear accents in four Irish (Gaelic) dialects. *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 965-968).

[27] Dalton, M. & Ní Chasaide, A. (2005). Tonal alignment in Irish Dialects. *Language and Speech 48*(4), 441-464.

[28] Dorn, A., O'Reilly, M. & Ní Chasaide, A. (2011). Prosodic signalling of sentence mode in two varieties of Irish (Gaelic). In *Proceedings of the 17th International Congress of Phonetic Science* (pp. 611-614).

[29] Dorn, A. & Ní Chasaide, A. (2015). Sentence mode differentiation in four Donegal Irish varieties. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

[30] O'Reilly, M. & Ní Chasaide, A. (2016). Modelling the timing and scaling of nuclear pitch accents of Connaught and Ulster Irish with the Fujisaki model of intonation. In *Proceedings of the 8th International Conference on Speech Prosody* (pp. 355-359).

[31] Dalton, M. & Ní Chasaide, A. (2007). Melodic alignment and micro-dialect variation in Connaught Irish. In C. Gussenhoven & T. Riad (Eds.), *Tones and Tunes: Studies in Word and Sentence Prosody* (Vol. 2). Mouton de Gruyter, Berlin, pp. 293-315.

[32] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy, A., Barnes, E. & Gobl, C. (2019). Leveraging phonetic and speech research for Irish language revitalisation and maintenance. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 994-998).

[33] Congregation of Christian Brothers. *Graiméar Gaeilge na mBráithre Críostaí.* (1960, 1999). Dublin: An Gúm.

[34] Congregation of Christian Brothers. (1988). *New Irish Grammar*. Dublin: C.J. Fallon.

[35] Uí Dhonnchadha, E., Nic Pháidín, C. & van Genabith, J. (2005). Design, Implementation and Evaluation of an Inflectional Morphology Finite-State Transducer for Irish. *Machine Translation 18*(3), 173-193.

[36] Rannóg an Aistriúcháin. (1958). *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil.* Dublin: Oifig an tSoláthair.

[37] Carnie, A. & Guilfoyle, E. (Eds). (2000). *The Syntax of Verb Initial Languages.* Oxford University Press.

[38] Duffield, N. (1995). *Particles and Projections in Irish Syntax.* Dordrecht: Kluwer.

[39] McCloskey, J. (1983). A VP in a VSO language?. In G. Gazdar, E. Klein, and G. Pullum (Eds), *Order, Concord and Constituency.* Dordrecht: Foris, pp. 9-55.

[40] Ó Donnchadha, G. (2010). *Syntactic Structure Building and the Verbal Noun in Modern Irish: A Minimalist Approach.* [Ph.D. thesis]. University College Dublin.

[41] Stenson, N. (1981). Studies in Irish Syntax. In W. Werner (Ed.) *Ars Linguistica.* Tübingen: Gunter Narr Verlag.

[42] Nolan, B. (2012). *The Structure of Modern Irish: A functional approach*. Sheffield: Equinox.

[43] Bresnan, J., Asudeh, A., Toivonen, I. & S. Wechsler. (2016). *Lexical Functional Syntax*. Oxford: Wiley Blackwell.

[44] Pollard, C. & Sag, I.A.. (1994). *Head-Driven Phrase Structure Grammar.* University of Chicago Press.

[45] Asudeh, A. (2002). The Syntax of Preverbal Particles and Adjunction in Irish. *LFG02 Conference*. Athens: CSLI Publications.

[46] Sulger, S. (2009). *Irish Clefting - The LFG perspective*. [Ph.D. thesis]. University of Konstanz.

[47] Mel'cuk, I.A. (1988). *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

[48] Harrington, S., Singleton, D. & Henry, A. (2000). At the sharp end of language revival: English-speaking parents raising Irish-speaking children. *Centre for Language & Communication Studies Occasional Paper 57*, 1-12. Trinity College, Dublin.

[49] Hickey, T. (2012). ILARSP: A Grammatical Profile of Irish. In M. Ball, D. Crystal & P. Fletcher (Eds), *Assessing Grammar: The Languages of LARSP*. Bristol: Multilingual Matters.

[50] Goodluck, H., Guilfoyle, E. & Harrington, S. (2001). Acquiring Subject and Object Relatives: Evidence from Irish. In J. Kallen (Ed.), *Journal of Celtic Language Learning, Vol 6: First Language Learning*.

[51] Brennan, S. (2004). *Na chéad chéimeanna: luathfhorbairt Gaeilge mar phríomhtheanga: díriú ar an bhfóineolaíocht* [*First steps: early development of Irish as a primary language: focus on phonology*]. Western Health Board.

[52] Ní Ghloinn, A., E. Uí Dhonnchadha, and A. O'Keeffe. (2018). The design and annotation of the TEG learner corpus of Irish. *Proceedings of the Inter-Varietal Applied Corpus Studies International Biennial Conference*. Malta.

[53] Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. https://www.coe.int/en/web/common-european-framework-reference-languages.

[54] Council of Europe. (2011). *Common European Framework of Reference for Languages: Learning, teaching, assessment*.

[55] Council of Europe. (2018). *Common European Framework of Reference For Languages: Learning, Teaching, Assessment - Companion Volume with New Descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989.

[56] Ó Meachair, M.J. (2020). *The Creation and Complexity Analysis of a Corpus of Educational Materials in Irish (EduGA)*. [Ph.D. thesis]. Trinity College, Dublin.

[57] Seymour, P., Aro, M., & Erskine, J. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology 94*, 143–17.

[58] Stenson, N., and Hickey, T. (2014). In defense of decoding. *Journal of Celtic Language Learning 18*, 11-40.

[59] Stenson, N. & Hickey, T. (2018). *Understanding Irish Spelling*. https://www.cogg.ie/wp-content/uploads/Understanding-Irish-Spelling-A-Handbook-for-Teachers-and-Learners-by-Dr.-Nancy-Stenson-and-Dr.-Tina-Hickey-1.pdf

[60] Breacadh. (2022). *Mar a Déarfá*. Casla, Co. Galway. https://www.maradearfa.ie/

[61] Education and Library Board. (2011). *Fónaic na Gaeilge*. Belfast.

[62] Nic Pháidín, F. (2017). Assessing Irish literacy resources: guidelines for teacher and publishers. In *Exploring the Literacy Landscape: Proceedings of the 40th Conference of the Literacy Association of Ireland*.

[63] Barnes, E. (2017). *Dyslexia assessment and reading intervention for pupils in Irish medium education: Insights into current practice and consideration for improvement*. [M.Phil. thesis]. Trinity College, Dublin.

[64] Barrett, M. (2016). *Doras feasa fiafraí: Exploring special educational needs provision and practices across Gaelscoileanna and Gaeltacht primary schools in the Republic of Ireland*. [M.A. thesis]. University College Dublin.

[65] An Roinn Oideachais agus Scileanna. (2016). *Polasaí don Oideachas Gaeltachta 2017-2022*. https://www.education.ie/ga/Foilseach%C3%A1in/Tuarasc%C3%A1lacha-Beartais/Polasai-don-Oideachas-Gaeltachta-2017-2022.pdf.

[66] Barnes, E., Ní Chasaide, A. & Ní Chiaráin, N. (2020). Bilingual phonological awareness: when interdependence becomes interference. *Proceedings of the 15th Congress of the International Association for the Study of Child Language*.

[67] Barnes, E., Ní Chasaide, A. & Ní Chiaráin, N. (2018). The design and pre-testing of literacy and cognitive tasks in Irish and English. *Proceedings of the Literacy Association of Ireland 42nd International Conference* (pp. 1-11).

[68] Barnes, E., Ní Chiaráin, N. & Ní Chasaide, A. (2017). Departures from the "norm": how the phonology, morphology and orthography of the Irish language impact on literacy instruction and acquisition. In *Exploring the Literacy Landscape: Celebrating 40 Years of Research and Practice. Literacy Association of Ireland 40th Annual Conference* (pp. 22-32).

[69] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy, A., Barnes, E. & Gobl, C. (2019). Leveraging phonetic and speech research for Irish language revitalisation and maintenance. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, (pp.

Chapter 3
Corpora

3.1 What is a corpus?

A corpus is a collection of language samples in digital format that show how a language is used in practice. These language samples can take the form of collections of texts, recordings of spoken language and transcripts (phonetic or orthographic) of audio or video material.

Some corpora consist of large general purpose collections. These tend to be drawn from naturally occurring, pre-existing language samples whenever possible, but in many cases the corpora need to be created for specific purposes. Examples include speech corpora designed for speech synthesis, learner corpora designed to investigate the types of errors and other features of learner language or social media corpora designed to provide insight into online use of language.

3.2 Why are corpora important and for whom?

Corpora are essential for speech and language technologies as they provide knowledge of the conventional use of language, which is needed for both interpreting and generating written and spoken language. Corpora are needed to build the resources, knowledge bases and models of Part I. The empirical evidence gleaned from a corpus shapes the linguistic information that we find in dictionaries, grammar books and thesauri. Corpora show language in context, allow the comparative frequency of both words and structures to be calculated and provide valuable information on every aspect of language structure - phonetic-phonological, syntactic, semantic and pragmatic.

Beyond the provision of resources, corpora are directly used in constructing and testing the core technologies of Part II, e.g. speech synthesis, speech recognition and machine translation. In some cases a corpus is needed as an inherent component of the system at run time (e.g. unit selection synthesis (Chapter 8) and machine translation (Chapter 11)). In today's world of big data and statistically-driven or neural based technologies deploying deep neural nets, the availability of large corpora is essential. For many of the applications envisaged in Part III, corpora also play an important role. For example, information from learner corpora, coupled with linguistic knowledge (derived from other types of corpora), provides essential input into state-of-the-art educational CALL applications.

In recent years, the most significant developments and progress made in the field of NLP (Chapter 5) have been based on the use of word embeddings and transformer models. These are a type of neural-network approach that more accurately learns about the nature of language by considering the context of words. The first transformer model for Irish – gaBERT [1] – trained upon the limited amount of existing corpora – is already showing an impact on the quality of existing NLP tools. The close link between corpora availability and the quality of language technology tools is therefore becoming more evident, and as such should be of primary focus for Irish going forward.

The public thus stand to benefit from corpora in many ways. The principal benefit will derive indirectly from the speech and language technologies themselves, even where the corpus contents are not visible to the user. In certain cases, corpora will be hosted with searchable interfaces on public websites, which users can directly access for their own purposes.

3.3 Types of corpora

A corpus may be composed of written or spoken language samples or both. These language samples are stored on computers as text files and audio/video files.

Written corpora typically consist of language samples taken from edited published sources such as books, newspaper articles, reports, manuals and textbooks. A wide range of topics and genres are usually included for language representativeness. These written corpora tend to be grammatical and use standard spelling, and therefore provide an idealised model of the language. More recently, due to increased use of the internet and social media, there is a need for language applications to be able to also handle user-generated content (UGC) e.g. from social media platforms such as Twitter or Facebook, which tends to be less grammatical and may contain non-standard spellings and abbreviations that pose challenges for processing tools. UGC provides a snapshot of the language in everyday social media use and such corpora are used for both sociolinguistic and NLP purposes. Written corpora not only appear in raw text format but can also contain useful linguistic annotations that can be carried out manually, semi-automatically or automatically (see § 3.4). These corpora are particularly valuable in the development of tools that automatically process human language (see Chapter 5).

Spoken corpora with orthographic or phonetic transcriptions can provide essential additional information to enhance what can be gleaned from text-only corpora. These are composed of broad general purpose collections of audio/video language samples, spontaneous or scripted, depending on the purpose of the corpus. Spontaneous language samples are used to document phenomena that are less typical in written language, e.g. discourse features, certain lexical items, hesitations, repetitions, reformulations, incomplete sentences etc. The orthographic/phonetic transcriptions are synchronised with the audio file for enhanced searching and further speech/language processing. For representativeness, spontaneous spoken corpora tend to contain a wide variety of speakers in a wide variety of naturalistic settings. Spoken corpora are essential for language documentation and language learning applications.

Speech corpora targeting speech technology are very diverse, ranging from such general-purpose spoken corpora (of spontaneous speech or read speech) to semi-spontaneous speech (where the speech is spontaneous but the topic constrained) to the reading of very carefully designed materials. As there is no single spoken ‘standard’ form of Irish, it is important that corpora are provided for the various dialects. While speech corpora are primarily acoustic (audio) recordings, articulatory or aerodynamic signals may be included as appropriate. The diversity of the corpora reflect the diversity of the technologies they are intended for. Therefore, the different types of corpora are discussed in the context of these technologies in Chapters 8, 9, and 10 and some main features are summarised here.

For speech synthesis (Chapter 8) both the quality of the speaker’s voice and the quality of the audio recording are of paramount importance. ‘Clean’ recordings are made in a high-spec studio, typically of prepared materials read by an individual (carefully chosen) subject. For speech recognition, in addition to very large general corpora, which include many voices, dialects, and recording conditions, one needs specific corpora that match the envisaged users and applications for which the system is intended, e.g. children in educational settings (see Chapter 9). Similarly, effective dialogue systems (Chapter 10) require conversational speech corpora to allow the modelling of real-life dialogues.For the linguistic analysis and modelling of Irish prosody and articulation (see Chapters 2.4 and 7), carefully designed corpora are needed that illuminate the precise features of interest.

3.4 Corpus preparation and annotation

The first task in creating a corpus after the corpus design has been finalised is to collect the language samples. Corpora are then often annotated or labelled with linguistic information. In the case of spoken audio files, the recordings may need to be transcribed, at least orthographically (but see more below).

3.4.1 Written corpora

In the case of written texts that are already in digital format, they usually require extensive ‘cleaning’ to remove unwanted material. A corpus of plain text is very useful but is usually only a first step. For most purposes the information of interest to an application is explicitly annotated (labelled, tagged) in the data. Both structural and linguistic annotations are commonly applied. Figure 3.1 shows an example of plain text where structural annotations for paragraph <p> and sentence <s> structure are explicitly shown in the data: *Bhí Máire fuar* ‘Mary was cold’. *Chuir sí geansaí olla uirthi* ‘She put a woollen jumper on’.

```
<p>
<s> Bhí Máire fuar. </s><s> Chuir sí geansaí olla uirthi. </s>
</p>
```

Figure 3.1: Plain text with paragraph and sentence structural mark-up

Figure 3.2 shows a short sentence which includes structural and linguistic annotation. In this example, each word is explicitly annotated with its lemma (dictionary headword) and a tag showing part-of-speech (verb, noun etc.) and grammatical function information (subject, object, etc.) as well as head-dependency relations (e.g. #4->3; token 4 is dependent on token 3). Such labelling is fundamental to most tools and applications described in Chapter 5.

```
<p><s>
<w lemma=”cuir” tag=”Verb+PastInd@FMV#1->1”>Chuir</w>
<w lemma=”Máire” tag=”Prop+Noun+Fem+Com+Sg@SUBJ#2->1”>Máire</w>
<w lemma=”geansaí” tag=”Noun+Masc+Com+Sg@OBJ#3->1”>geansaí</w>
<w lemma=”olann” tag=”Noun+Fem+Gen+Sg@N<#4->3<”>olla</w>
<w lemma=”ar” tag=”Pron+Prep+3P+Sg+Fem@PP_OBL#5->1”>uirthi</w>
<w lemma=”.” tag=”Punct+Final#6->6”>.</w>
</s></p>
```

Figure 3.2: A sentence annotated with lemma, part-of-speech, grammatical function and head-dependency relation for ‘Chuir Máire geansaí olla uirthi’

A corpus of text that has been annotated with information regarding the syntactic structure of the text is known as a Treebank (Figure 3.3). Such a corpus is referred to as a treebank because sentences are represented as parse trees, where words and phrases are connected through branch-like links that describe grammatical relationships and structures (see Appendix A). Treebanks are needed both for syntactic linguistic research and to train statistically based syntactic parsers (see Chapter 5).

```
# sent_id = 904
# text = Creidtear gur go mailíseach a tosaíodh an tine.
1 Creidtear   creid            VERB    VTI      Mood=Ind|Tense=Pres|Voice=Auto               0    root
2 gur         is               AUX     Cop      Tense=Pres|VerbForm=Cop                      4    cop
3 go          go               PART    Ad       PartType=Ad                                  4    mark:prt
4 mailíseach  mailíseach ADJ   Adj     Degree=Pos                                            1    ccomp
5 a           a                PART    Vb       PartType=Vb|PronType=Rel                     6    mark:prt
6 tosaíodh    tosaigh          VERB    VTI      Mood=Ind|Tense=Past|Voice=Auto               4    csubj:cleft
7 an          an               DET     Art      Definite=Def|Number=Sing|PronType=Art        8    det
8 tine        tine             NOUN    Noun     Case=NomAcc|Gender=Fem|Number=Sing           6    obj
9 .           .                PUNCT   .        _                                            1    punct
```

Figure 3.3: A sentence annotated in the Irish Universal Dependency Treebank, containing: sentence ID number, raw text, lemma, POS, morphological features, along with grammatical structure labels and attachments

Semantically annotated corpora are corpora that contain semantic annotations regarding individual lexical items and grammatical structures illustrated in Figure 3.4. A corpus in which arguments of verbs and other predicates are annotated is known as a proposition bank, e.g. PropBank [2] (see Chapter 2, semantics). All verbs in a selected corpus are annotated, which gives wide coverage of verb usage and variability. Named entities can also be labelled which helps to distinguish types of noun phrases and determine their semantic roles. This information is necessary for downstream tasks such as quality information retrieval, text summarisation and machine translation.

```
[Mr.  Bush] met [him] [privately], [in the White House], [on Thursday].
 ARG0      REL ARG1  ARGM-MNR    ARGM-LOC          ARGM-TMP
```

Figure 3.4: A semantically annotated sentence based on an example in English PropBank Annotation Guidelines [3] using Propbank semantic labelling ARG0: agent, ARG1: patient, ARGM-MNR: manner modifier, ARGM-LOC: locative modifier and ARGM-TMP: temporal modifier.

3.4.2 Speech corpora

In the case of speech corpora intended for speech technology, the individual sounds in an audio file are identified and the signal segmented to identify where they begin and end. Checking of annotation involves visual inspection of the speech spectrogram (upper part of Figure 3.5) and careful listening, to ensure that the transcription matches what was spoken, that segmentations are correct, that hesitations and speech errors are identified, etc.

Further levels of annotation may follow, depending on the purpose for which the corpus is intended. For example, if prosodic information is needed (for implementation in synthesis/dialogue systems) the pitch contour would be shown and important points notated. When such additional annotation is included, the corpus becomes a rich database – useful for phonetic-linguistic analysis (Chapter 2.4 and for the modelling of Irish speech processes for technology (Chapter 7).

3.4.3 Manual and automatic segmentation

Corpora can contain many millions of words or many hours of recorded speech. Typically, the larger the corpus, the more useful it is. Corpus annotation involves a considerable investment of time and expertise. Annotated corpora that have been manually checked and verified (by two or more experts) are referred to as gold-standard corpora and are directly used in the training and evaluation of speech/language processing tools and in the building of technologies. At the initial stages of corpus preparation, linguistic annotations are typically carried out by hand, a process that demands knowledge of the language (or dialect) and of the linguistic aspects being included. However, annotating large corpora by hand is an impractical task. If a sufficiently large body of hand-annotated data is available, this data can be used, through machine learning, to automate (or semi-automate) the process of annotation for additional corpora (this method is sometimes referred to as bootstrapping). Gold-standard corpora can then be developed more quickly through manual correction of automated labelling, as opposed to manually annotating data from scratch. For speech corpora, automatic segmentation algorithms can be used to provide an initial alignment of the individual sounds (phonetic description) to the speech waveform.

3.5 Specialised corpora

This section illustrates some examples of specialised corpora, which are of particular relevance to the Digital Plan.

Parallel corpora contain both original texts and their translations, which are aligned at the sentence level. Both statistical machine translation (SMT) and neural machine translation (NMT) learn patterns from large collections of parallel text in order to predict translations. For example, an English-Irish parallel corpus contains a large body of English text aligned with its exact translated equivalent in Irish. Sourcing good parallel data can be a challenge, particularly in the case of lesser-resourced languages such as Irish. Often the main task of creating a parallel corpus involves the identification and collection of previously translated data from translators or translation bodies

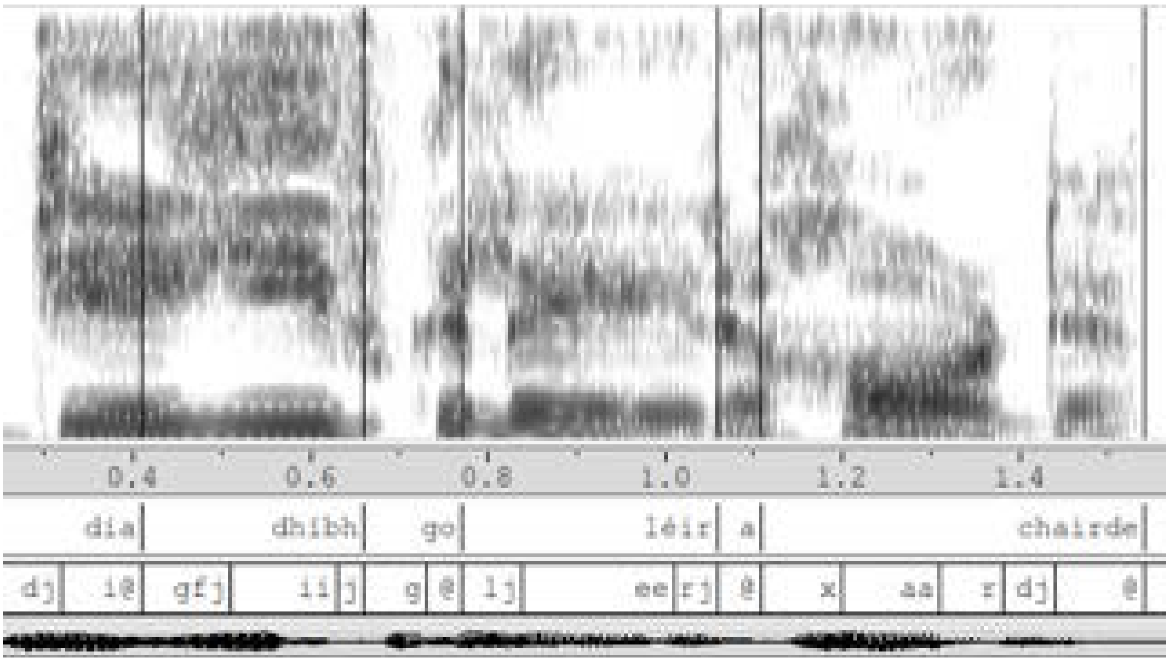


Figure 3.5: A sample from the ABAIR annotated speech corpus, showing the speech spectrogram with orthographic and phonetic segmentation and transcription.

or contacting appropriate government departments or semi-state organisations. In addition, software tools called 'web-crawlers' are sometimes used to find parallel texts on the internet. Ensuring correct alignment across sentences to prevent the SMT system learning incorrect translations is vitally important.

Comparable corpora can also be useful for machine translation. In contrast to parallel corpora, these are data sets with text in English and a rough equivalent in Irish. This may arise from the generation of content on a topic in both languages simultaneously as opposed to translating from one to another. An example of this is Wikipedia (see Appendix A) and online news sites, where content articles are often created independently, yet cover the same topic. Comparable corpora can provide a rich insight into the idiomatic differences between languages, which is valuable for both linguistic analysis and machine learning.

Learner corpora contain samples of learner language collected at different stages in the learning process. These language samples can include written and spoken language. Learner corpora are an essential resource in seeking to understand the language-learning process. They also allow us to identify specific aspects of language that present particular difficulties for the learner, e.g. aspects which are inherently complex, rare in general usage or very different from the learner's first language. The empirical evidence, concerning the order and rate at which specific elements are mastered and the levels of difficulty they present, are valuable for fine-tuning of curricula and language-learning materials. This also provides essential underpinning to the development of CALL technologies which target specific areas of difficulty, and which are appropriate to the various stages of acquisition.

Domain-specific corpora are collections of text gathered according to specialised domains or genres, such as legal text, scientific materials, educational materials, newswire, and so on. Statistically-based language-processing tools that deal with the varying nature of text rely on the availability of such domain-specific text for their development and testing. Gathering of such specific data requires a targeted collection effort. However, newswire and online content (where licensing permits) can be acquired through web-crawling.

Extensive speech corpora are needed for all speech technology development. The size and composition of the corpus depends on the synthesis methodology to be adopted (see Chapter 7), but regardless of size, phonetic coverage is needed to ensure that all the sounds of the dialect are included in all contexts they occur in. A very large corpus tends to provide good coverage, but is costly in terms of the manpower needed for annotation, correction, etc.

3.6 What has been done to date?

Corpora developed for Irish to date, or currently under development are outlined in Tables 3.1 and 3.2. By international standards, these are all relatively small in size and in all cases further materials need to be collected or recorded. In addition, new and enhanced (semi-)automatic annotation methods need to be initiated. Further details of some of these corpora can be found in Appendix A along with a summary of recommendations for future development. All of these resources and more are listed in the European Language Grid [4], which will continue to be updated with new resources on a regular basis.

Table 3.1: Written corpora

The availability of written corpora is fundamental to the development of today's state-of-the art technologies. Restrictive licensing on existing corpora holds many such developments back, however. It is worth noting therefore, that in addition to the corpora listed in the table below, the development of the National Corpus of Ireland has been under way since January 2022 and is being carried out by researchers at Fiontar & Scoil na Gaeilge in DCU and TCD. This work is a significant step towards complementing existing corpora with language material from the last 20 years and addressing corpus accessibility issues.

Most large-scale corpus initiatives to date have involved text corpora (Table 3.1). The collection of speech corpora for speech technology began in relatively recent years. Some examples of speech and spoken corpora are included in Table 3.2.

Corpus	Description	Annotation Details
1. NCI Irish Corpus http://corpas.focloir.ie/	30 million words of written text designed to be balanced and representative	Automated rule-based tokenisation and POS tagging.
2. Gold Standard POS tagged corpus	3,000 sentences randomly selected from corpus (POS-tagging development).	Automatic rule-based tagging, with manual verification and correction
3. Irish Dependency Treebank (IDT) https://github.com/tlynn747/IrishDependencyTreebank Irish Universal Dependency Treebank (IUDT) http://universaldependencies.org/#ga	1020 parsed sentences of written language (gold standard). 4910 parsed sentences of written language (gold standard). Development ongoing.	Based on LFG-inspired annotation scheme. Automatically parsed, followed by manual checking and correction. UD version: IDT mapped to IUDT annotation scheme
4. Irish Crúbadán Web Corpus http://crubadan.org/languages/ga	100+ million words of web-crawled written data – updated daily	Automatically crawled online and automatically aligned.
5. Vicipéid https://ga.wikipedia.org	57K articles – Irish Wikipedia (online encyclopedia)	Structured linked data (text and images). Created manually by the public – reviewed by Wikipedia editors.
6. Gaois Corpus of Contemporary Irish http://www.gaois.ie/g3m/ga/	24.7 million words of contemporary text published from the beginning of the 21st century onwards	Manual and semi-automated curation.
7. Gaois Parallel Corpus https://www.gaois.ie/crp/ga/	Parallel corpus of aligned text segments from Irish and European Union legislation. 26 million Irish words and 24.5 million English words.	Manual and semi-automated curation.
8. Corpora of Social Media Text https://github.com/tlynn747/IrishTwitterPOS https://github.com/UniversalDependencies/UD_Irish-TwittIrish	1,500 Irish tweets lemmatised and POS-tagged TwittIrish Universal Dependency Treebank of over 800 Irish tweets	Automatically pre-tagged with standard POS-tagger. Mapped to Twitter POS tagset, parsed and manually verified Mapped to UD POS tagset, tagged for code-switching, parsed through bootstrapping IUDT and manually verified.
9. Corpus of Irish multi-word expressions https://gitlab.com/parseme/parseme_corpus_ga	1,700 Irish sentences tagged with multiword expressions	Manually annotated and fully reviewed according to PARSEME guidelines.
10. EduGA: Corpus of Educational Materials	7.5 million words from educational materials for teaching Irish and other subjects taught through the medium of Irish.	Automatic rule-based POS tagging with manual correction. Automatic sentence tokenisation with manual correction.
11. TEG Learner Corpus of Irish (Corpas Foghlaimeora TEG)	Approximately 200,000 words of written and spoken learner-generated language taken from Teastas Eorpach na Gaeilge (TEG) proficiency tests at intermediate to advanced levels..	Manual error correction/ mark-up followed by customized automatic rule-based POS tagging.
12. Irish-EU English-Irish Parallel Corpus European Language Resource Coordination (ELRC) http://www.lr-coordination.eu/resources eSTÓR https://estor.ie	Ongoing national translation data collection by DCU through ELRC-SHARE/ eSTÓR. Collected from various public bodies and government departments under varying licencing agreements -- mostly Open Data (PSI). 195,000+ parallel sentences Data collected for the purposes of Machine Translation development at national and EU level. Will grow with eSTÓR contributions.	English and Irish translated texts generated by professional translators, manually and semi-automatically aligned at the sentence level.
13. National Folklore Collection - Schools' Collection (Irish) http://Duchas.ie	Over 3.8 million words of Irish from the 1930s	Transcribed via crowdsourcing
14. Corpas Stairiúil na Gaeilge 1600-1926 http://corpas.ria.ie https://www.ria.ie/research-projects/focloir-stairiuil-na-gaeilge	90 million words in over 3000 texts published in Irish between 1600 and 1926. This corpus is the basis for the RIA Foclóir Stairiúil na Gaeilge.	Manual and automatic spelling standardisation followed by customized automatic rule-based POS tagging.
15. Corpas Filíocht shiollach na Gaeilge, circa 1200-1650 https://app.sketchengine.eu/#open	1+ million words of Classical Modern Irish, of which 300K words are freely available online	Manual and automatic spelling standardisation followed by customized automatic rule-based POS tagging.

Table 3.2: Speech and Spoken Corpora

Corpus	Description	Annotation Details
1. ABAIR General Speech Synthesis Corpus: www.abair.ie	Ongoing: currently 3 main dialects, total c. 25 hours of recorded speech	Parts segmented and aligned phonetic annotation (X-SAMPA, IPA); stress marking
2. ABAIR Compact Speech Synthesis An Corpas Beag	Ongoing: (c.3K prompts) scripted prompts for maximal phonetic coverage with minimal material. Currently designed for Munster and Connaught Irish. Work proceeding on other dialects .	Segmented and aligned phonetic annotation (X-SAMPA, IPA); stress marking
3. Comhrá Spoken Corpus *https://www.scss.tcd.ie/~uidhonne/comhra/*	240K words of transcribed spontaneous spoken language from all of the major of the dialects	Transcripts segmented and aligned with audio files. Automatic rule-based POS tagging.
4. An Scéalaí - Corpus Cliste: learner corpora for speech technology development and iCALL research	Under construction: data being gathered from senior primary, secondary, tertiary levels as well as adult learners	Not yet annotated
5. ABAIR MíleGlór Speech Recognition Corpus	Under construction: being collected using crowdsourcing and field recordings in Gaeltacht locations	Editing and refinements of corpus ongoing
6. TEG Learner Corpus of Irish (Corpas Foghlaimeora TEG)	Approximately 200,000 words of written and spoken learner-generated language taken from Teastas Eorpach na Gaeilge (TEG) proficiency tests at intermediate to advanced levels.	Manual error correction/ mark-up followed by customized automatic rule-based POS tagging.
7. ICCI International Comparable Corpas – Spoken and Written Irish	Under construction. Target: 600K words spoken language and 400K words written language.	Spoken data mainly requires new recordings. Written data being sourced.

3.7 Corpora – Recommendations

3.7.1 Continued data collection, maintenance and corpus creation

In the context of today's data-driven and statistically-based approaches, corpora that are larger and of higher quality enable the development of better language technology tools and applications. Therefore, while extensive new corpora are required for Irish, the corpora listed above need to be continuously expanded and curated. All aspects of future corpus development need careful planning, i.e. in terms of number and size of samples, the speakers targeted and their linguistic background, the time period in question, language varieties, types and level of annotation etc., thus ensuring that the corpora are suited to the specific research and technology development envisaged. The approach to development of all future corpora should be to keep up to date with international state-of-the-art approaches, standards and practices.

Some specific recommendations are as follows:

- Gold-standard annotated corpora (manually verified) are essential for the development of high quality speech and language technologies and are a valuable resource for linguistic research. For example, existing gold-standard POS-tagged corpora and treebanks require significant expansion to facilitate robust and reliable parser development.
- Large-scale Speech Corpora will need to be designed, collected and processed to meet the needs of the core technologies: synthesis, recognition, and spoken dialogue systems (Part II). These will need to cater for the various Irish dialects. As discussed in Chapters 8, 9 and 10, although many corpora will entail wide coverage of speakers (age, gender, etc.), specific corpora (e.g. of children's voices) will also be needed for targeted applications.
- Semantically Annotated Corpora. As there is no semantically annotated corpus currently available for Irish, development of a semantic tagger by researchers in TCD and DCU has commenced as part of the National Corpus of Ireland project. Such a resource is a priority in order to provide semantically informed tools and applications, e.g. CALL, information retrieval, document summarisation and robust parsing. As semantic annotation is a detailed and time-consuming endeavour, priority for annotation could be given to high frequency verbs as was done for Basque [5]. For maximum efficiency, all available lexical resources (see Chapter 4) and semi-automated annotation methods (see Chapter 5) should be used to speed up development.
- Domain-specific corpora: as the demand for tailored NLP tools grows, we will see the demand for domain-specific corpora grow (e.g. corpora of scientific articles, educational materials, legal texts and other domain-specific texts)
- The NCI written corpus needs to be significantly

expanded to meet the needs of modern corpus-based lexicography. It also needs to include a significant amount of transcribed spoken language. As the NCI is of mixed domain, detailed metadata is important for enabling sentences to be selected or filtered out based on genre or source type. This type of large corpus is also vital for training word embeddings for NLP tools, language models for Statistical MT or predictive text and back-translation for synthetic data creation for Neural MT (Chapter 5).

- A number of types of spoken corpora are required to support speech and language technology, including domain-specific spoken corpora, which are necessary for identifying and quantifying specialised vocabulary and discourse features.
- An Vicipéid (Irish Wikipedia) needs to be continuously expanded as an ongoing source of freely available digital texts. As Wikipedia is a community-driven initiative, a nation-wide effort is required to promote the use of an Vicipéid, along with the training of new volunteer editors. Innovative approaches involving educational institutions can help ensure the continued growth of this resource.
- User-generated content (e.g. internet and social media data) will need to be collected, curated and analysed in order to maintain an up-to-date insight into how the language is used and evolving online. Sentiment analysis tools (assessing opinions shared online or in large scale documentation) also rely on the availability of such corpora.
- Various types of learner corpora (written and spoken) are vital for empirically driven design of language learning tools.

3.7.2 Corpus annotation, visualisation and verification tools

The annotation and classification of text and speech is a time-consuming and difficult task, which requires highly-skilled researchers. It involves close analysis, categorisation and labelling of phonemes, words, phrases, relationships between words, between constituents in a sentence or between referrers and referents in discourse. In speech corpora the identification and labelling of salient acoustic/articulatory features (e.g. pitch peaks) may also be needed. In order to produce large, high-quality corpora, it is recommended that annotation, visualisation and verification tools be developed that can make these annotations feasible on a large scale, thereby also facilitating analytic research. Such tools should also allow for easy labelling and linking/unlinking of elements.

Chapter 13 describes how crowdsourcing can be effective for scaling certain kinds of sub-tasks that do not require specialist skills. In the broad field of NLP, crowdsourcing platforms such as Amazon Mechanical Turk[1] and Appen[2] are widely used methods for engaging the public in data annotation. However, national platforms designed specifically for Irish speakers are required to facilitate this approach to ensure the relevant Irish skills are found, e.g Meitheal dúchas.ie.

3.7.3 Corpus query and analysis platforms

In addition to the development of corpora, it is important that the annotated data in these resources can be easily accessed, searched and analysed. Therefore, there is a clear need for suitable platforms that users (of all technical ability) can use to browse or query the resources (e.g. the POS-tagged corpora, treebanks, phonetically annotated corpora, transcripts, audio files etc.). These platforms should also allow integration with existing processing and analytical tools. It is recommended that (i) the suitability of current open-source tools be investigated and (ii) new language-tailored tools be developed where necessary.

References

[1] Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Ó Meachair, M.J. & Foster, J. (2022). gaBERT – an Irish Language Model. In *Proceedings of the 13th Conference on Language Resources and Evaluation* (pp. 4774-4788).

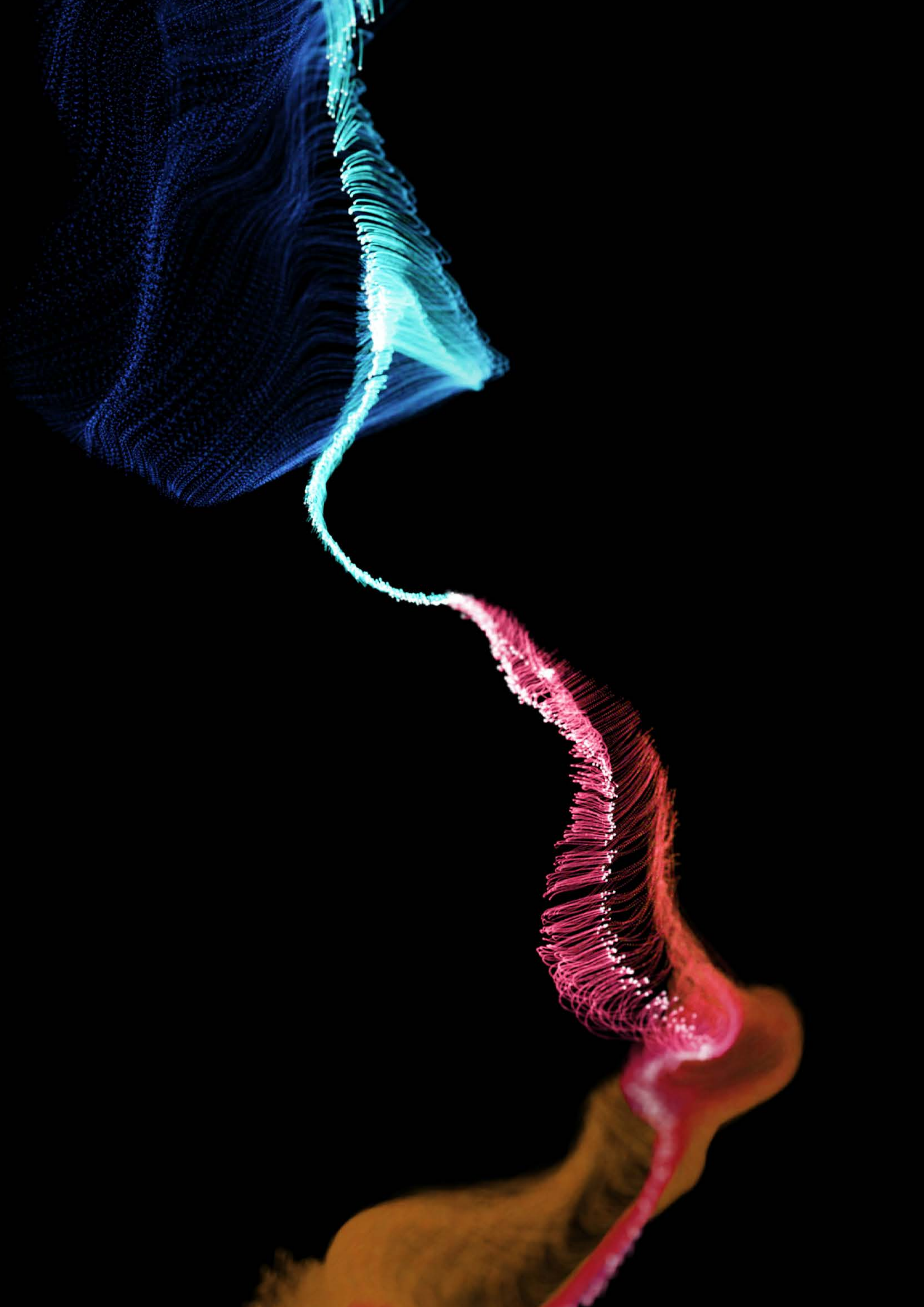
[2] Palmer, M., Kingsbury, P. & Gildea, D. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1), 71-105.

[3] Bonial, C., Bonn, J., Conger, K., Hwang, J. Palmer, M. & Reese, N. (2015). English PropBank Annotation Guidelines. University of Colorado.

[4] ELG Consortium. (2022). *European Language Grid*. https://live.european-language-grid.eu/

[5] Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Pociello, E. & Quintian, M. (2006). Improving the Basque WordNet by corpus annotation. In *Proceedings of the 3rd International WordNet Conference* (pp. 287-290).

[1] https://www.mturk.com
[2] https://appen.com



Chapter 4
Linguistic Knowledge Bases

4.1 What is a linguistic knowledge base?

Linguistic knowledge bases are organised collections of language data. Electronic dictionaries, terminology databases, thesauri, gazetteers (people and places) and glossaries are examples of such linguistic resources. Knowledge bases contain information that can range from translation and meaning, to grammatical information, and sometimes details of usage in context. They can contain linguistic information about various elements of the language, such as individual words, multi-word entities and the meaningful links between them. While corpora (Chapter 3) contain naturally occurring sentences and utterances, they may not contain every possible form of the word. Linguistic knowledge bases, on the other hand, endeavour to hold complete inventories/lists of word forms and how they can be used. In this way, corpora and knowledge bases provide complementary information about a language; corpora provide information about conventional usage and most frequent forms, while knowledge bases provide information about potential uses and restrictions.

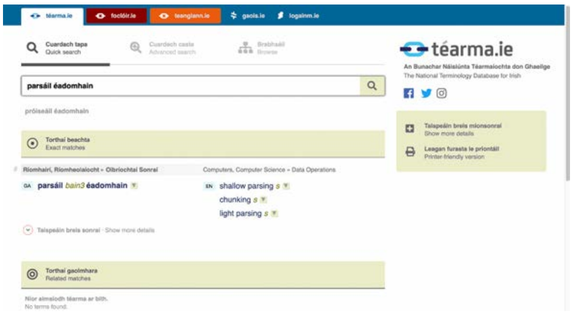
In this chapter, we describe three main categories of knowledge base: lexical, syntactic and semantic knowledge bases. Lexical knowledge bases contain information at the word and multi-word level. Syntactic knowledge bases contain information about how words can combine to form grammatical sentences. Semantic knowledge bases contain information regarding the meanings of individual words or combinations of words.

4.2 Why are linguistic knowledge bases important?

As with corpora, knowledge bases are essential in language technology research and their value is highlighted in most chapters of this document. At the very basic level, for spellchecking, the existence of a word in a language can be verified by computer by checking a lexical database. A lexical database can also be used in an interactive tool that allows online users to hover over or click on a word unknown to them in order to access a dictionary meaning entry or translation (see Chapter 13).

Bilingual terminology databases are resources that are of particular use in finding the correct terminology to use in translations. These shared repositories are often used by a team of translators to ensure terminology consistency in a technically supported translation environment (see Chapter 11), particularly for specific domains, such the legal, medical, educational domains etc. Figure 1 shows a search for an entry for the language technology domain. Terminology databases are also valuable for the development of semantic analysis tools (Chapter 5) and language-learning tools (see Chapter 14).

There is also a need for more monolingual dictionaries for Irish. These are dictionaries that provide a description of meaning and grammatical use for Irish words through the medium of Irish. If a dictionary user is checking the entry for an Irish word, while learning or working through the medium of Irish, they should not be expected to switch to using English while accessing a lexical resource. This is especially true for pupils of Irish medium schools.



parsáil éadomhain

próiseáil éadomhain

Ríomhairí, Ríomheolaíocht » Oibríochtaí Sonraí | Computers, Computer Science » Data Operations

GA parsáil *bain3* éadomhain

EN shallow parsing *s*
chunking *s*
light parsing *s*

Níor aimsíodh téarma ar bith.
No terms found.

Figure 4.1: Terminology search on téarma.ie

In order to perform grammar checking on a document, computers need to know about the grammatical rules of a language. This knowledge is essential for accurate understanding and generation of language. For example, *Thug an stiúrthóir doiciméad* 'The director gave a document' is not a grammatical sentence because one of the essential participants of the verb 'to give' is missing, i.e. the recipient. A rule-based grammar checker could access such information in order to determine whether a sentence violates any of the grammatical restrictions. Syntactic knowledge about word order, word combinations and grammatical relationships can be encoded in computational grammars. In addition to grammar checking (see Chapter 13), these types of knowledgebase are fundamental to many language processing tasks (see Chapter 5), and applications such as language generation (see Chapter 6), information extraction (see Chapter 12) and language learning applications (see Chapter 14).

From a meaning perspective, semantic knowledge is important for many language processing tasks. For example, a rule-based machine translation engine translating from Irish to English needs some way to distinguish the senses of *fiach* ('raven', 'debt', or 'hunt') in order to provide the correct English translation in context. Being able to resolve such ambiguities can also improve the performance of search engines (Chapter 12) by enabling the indexing of documents according to word senses rather than just word spelling occurrences.

In addition to monolingual semantic databases, in a multilingual context, e.g. for machine translation, it becomes necessary to link semantic databases between languages. Because of their importance in Natural Language Processing (NLP), semantic databases and networks have been developed for a large number of languages, and many have been linked together. This enriches terminology databases, allows for the development of thesauri, and assists in NLP tasks such as named entity recognition (NER) (Chapter 5), machine translation (Chapter 11), cross-lingual information retrieval (Chapter 12), among others.

DBpedia is a type of knowledge base that allows for better extraction of information from Wikipedia articles. The availability of a DBpedia for a language allows the Wikipedia pages for that language to be searched more accurately and efficiently. DBpedia, like Wikipedia, is a crowd sourcing community effort. DBpedia is divided into language chapters. Each language chapter is based on volunteer work (a crowd sourcing initiative) and is responsible for extracting information from different language versions of Wikipedia.

Linked Data is also a type of knowledge base. The goal of Linked Data is to represent data in a structured format that allows computers to read, query, extract and interpret the data. Each Linked Data dataset should also provide links to other datasets so that a computer programme can navigate through different domains and find additional relevant data. The availability of Open Linked Data (data which is released under an open licence) on the web enables the immediate use of the data by application developers.

4.3 How do they work?

Computers rely heavily on language resources that contain information about the lexical, syntactic and semantic restrictions and conventions of language use. In this section we give an overview of common lexical, syntactic and semantic language resources.

Lexical resources

Information about the individual words in a language is stored in lexical knowledge bases. They can take the form of databases, networks and other structured digital representations of data such as XML files. Lexical databases are often initially derived from machine-readable versions of published dictionaries, e.g. [1-3] form the basis of the www.teanglann.ie website resource.

Dictionary and terminology database entries usually contain headwords, e.g. *bord* rather than all forms including *bhord*, *bhoird*, *mbord*, etc. Word processing tools (see Chapter 13) can search for wordforms in a lexical database. However, from a computational efficiency perspective, NLP applications can often benefit from using morphology tools (see Chapter 5) that search through smaller networks that contain lemmas or headwords only, and the rules to generate the word forms associated with a headword.

In contrast to dictionaries, which tend to focus on general purpose language and individual words, terminology databases focus on domain-specific language. These databases often provide examples of terms in use, and the context in which they are most suitable.

Given that much knowledge and terminology is often coined in English first (and then translated into other languages), one possible strategy for keeping Irish language termbases up to date is to monolingually extract and identify new terms in newly published English technical/specialised text and then manually translate those terms into Irish. This way Irish terminology is ready to use for when documents in Irish need to be produced.

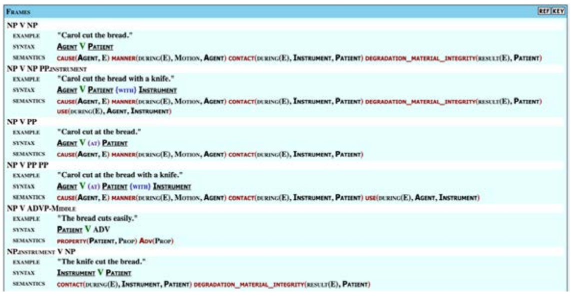
In addition, specifically tailored tools and methods [4] can be used to help with terminology extraction from scientific corpora and bilingual terminology extraction from translation memories.

Syntactic resources

Language technologies that processes language on a scale larger than the word or phrase level i.e. at the sentence or document level, require knowledge of the rules of syntax of the language. The automatic detection of syntactic structures in a language is known as syntactic parsing (see Chapter 5). Computational grammars encode representations of the syntactic structure of a language and are essential resources for rule-based parsing technology. Phrase structure grammars and dependency grammars contain information about the types of words (nouns, verbs, adjectives etc.) that can be used together and the orders that they are permitted.

A verb is usually the central component of a sentence and verbs often have requirements and restrictions as to what particular words can be used with them (e.g. nouns, prepositions etc.) and where they are positioned in a phrase in relation to the verb. For example, *chuaigh isteach sa rang an dalta (incorrect word order), *theip liom sa scrúdú (incorrect preposition), and so on, are ungrammatical constructions in Irish. Computational grammars make use of linguistic resources such as verb valency dictionaries which contain information about the number of participants associated with a verb, e.g. the verb tabhair 'give' usually requires three participants; the giver, the thing being given and the recipient. A valency dictionary can be used to create an annotated corpus such as PropBank[1] (see Chapter 3).

VerbNet [5, 6] encodes verb-argument structure, i.e. how many participants and of what type are needed by a particular verb. Figure 4.2 shows an example of how arguments for the English verb 'cut' are encoded in VerbNet. The development of a VerbNet-type resource for Irish would provide information in a computer-friendly way which is vital for grammar development, parsing and pedagogical purposes. It would also complement morphological analysis and integrate with other lexical and semantic resources for Irish.



Frames REF KEY

NP V NP
Example "Carol cut the bread."
Syntax Agent V Patient
Semantics cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient) degradation_material_integrity(result(E), Patient)

NP V NP PP.instrument
Example "Carol cut the bread with a knife."
Syntax Agent V Patient (with) Instrument
Semantics cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient) degradation_material_integrity(result(E), Patient) use(during(E), Agent, Instrument)

NP V PP
Example "Carol cut at the bread."
Syntax Agent V (at) Patient
Semantics cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient)

NP V PP PP
Example "Carol cut at the bread with a knife."
Syntax Agent V (at) Patient (with) Instrument
Semantics cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient) use(during(E), Agent, Instrument)

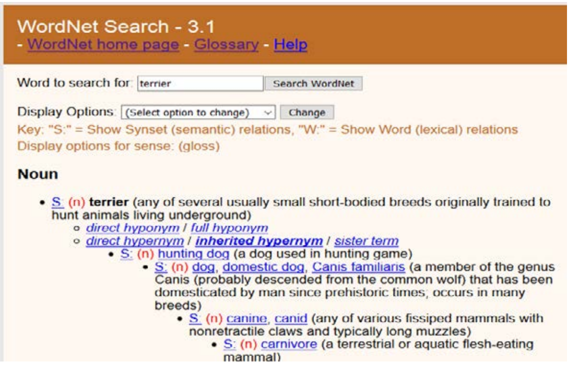
NP V ADVP-Middle
Example "The bread cuts easily."
Syntax Patient V ADV
Semantics property(Patient, Prop) Adv(Prop)

NP.instrument V NP
Example "The knife cut the bread."
Syntax Instrument V Patient
Semantics contact(during(E), Instrument, Patient) degradation_material_integrity(result(E), Patient)

Figure 4.2: Representation of the argument structure for 'cut' in the English VerbNet[2]

Semantic resources

One of the fundamental problems in getting computers to understand human language is that words can be ambiguous in meaning (e.g. the Irish word *fiach*, which can mean either 'a debt', 'a raven', or 'hunting'). Semantic databases provide a 'sense inventory' for a language, and also provide the means by which a computer can distinguish word senses in context. In general terms, from an artificial intelligence perspective, a semantic network encodes some of the 'real-world knowledge' that is required for computers to understand and process texts in a non-trivial way.



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: terrier [Search WordNet]

Display Options: (Select option to change) [Change]
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss)

Noun

- S: (n) **terrier** (any of several usually small short-bodied breeds originally trained to hunt animals living underground)
  - direct hyponym / full hyponym
  - direct hypernym / ***inherited hypernym*** / sister term
    - S: (n) hunting dog (a dog used in hunting game)
      - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds)
        - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
          - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal)

Figure 4.3: WordNet Online[3]

Semantic knowledge can be encoded as networks of relationships between words and concepts. Semantic networks are richer than traditional thesauri, which generally only record (near) synonyms and sometimes antonyms. These networks are usually known as wordnets, after Princeton's English-language Wordnet [7] which is a large lexical database of English nouns, verbs, adjectives and adverbs (see Figure 4.3).

Senses are grouped into sets of cognitive synonyms (synsets) based on meaning similarity, each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. See Figure 4.4 for an example of how the term 'word' is represented in the Princeton WordNet lexical resource.

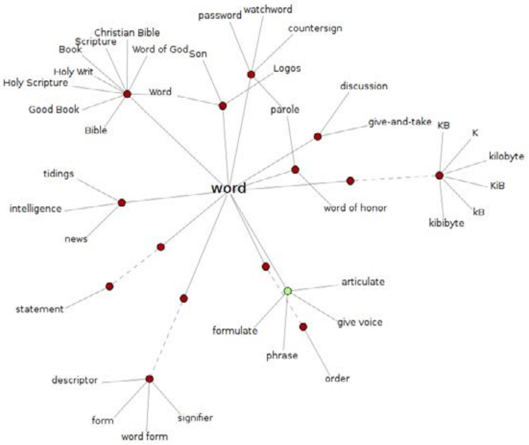


Figure 4.4: WordNet representation for the term 'word'

FrameNet [8] deals with sentences rather than words (i.e. in contrast to WordNet and VerbNet), and contains information not only on lexical semantics but also predicate-argument semantics. The use of FrameNet is a growing area of research in NLP, for example in the analysis of multiword expressions [9] and semantic parsing [10].

[1] https://propbank.github.io/
[2] https://verbs.colorado.edu/verbnet/
[3] http://wordnetweb.princeton.edu/perl/webwn

Knowledge graphs

In the field of knowledge representation and reasoning, a knowledge graph is a knowledge base that uses a graph-structured data model or topology to integrate data. Knowledge graphs are often used to store interlinked descriptions of entities – objects, events, situations or abstract concepts – while also encoding the semantics underlying the used terminology[1]. Knowledge graphs are closely linked to the Semantic Web and are used widely in search engines such as Google and Bing, in social networks such as LinkedIn and Facebook, recommender systems such as Amazon and question-answering systems such as Siri or Alexa. They are powerful NLP solutions that are often combined with machine-learning based systems in an approach known as Symbolic AI. Their strength comes from underlying datasets that have been labelled according to taxonomies and ontologies that help capture meaningful (semantic) relationships between entities (e.g. products, people, companies etc.).

4.4 What has been done for Irish?

In this section we give an overview of some of the main lexical, syntactic and semantic language resources which are under development for Irish.

Lexical resources

To date, the development of lexical resources for Irish has received continuous directed funding (e.g. through Foras na Gaeilge) and thus has demonstrated steady progress over the past several years. Current resources include terminology databases, bilingual dictionaries and a monolingual dictionary, and these are all available to the public either through websites or smart phone apps (foclóir.ie & teanglann.ie). In addition, the Irish national terminology database (téarma.ie) is accessible freely online and is used widely[2] by the general public, educational institutions, translators in public administration, in European institutions and by freelance translators. Table 4.1 provides a summary of the various lexical resources that are currently under development for the Irish language.

Resource Name	Description	Details
Foclóir.ie *www.foclóir.ie*	The New English-Irish Dictionary (Foras na Gaeilge)	• 50,000 (48,056) Headwords • 125,000 (123,309) Concepts
Teanglann Dictionary and Language Library *www.teanglann.ie*	(Foras na Gaeilge)	• Foclóir Gaeilge-Béarla (Ó Donaill, 1977): 42,106 • English-Irish Dictionary (de Bhaldraithe, 1959): 43,082 Headwords • An Foclóir Beag (monolingual) (Ó Dónaill & Ua Maoileoin, 1991): 16,753 Headwords • Morphology Database: 43,000 entries
Thesaurus *www.potafocal.com*	Irish language thesaurus based on the Irish WordNet (Líonra Séimeantach na Gaeilge)	• 30,000+ entries automatically generated
Gluais Tí Gluais Beo *www.potafocal.com*	Searchable database of bilingual glossaries Provides context of use, translations and information on gender, grammatical use, parts of speech and frequency of use	• Gluais Tí: 5,000 entries • Gluais Beo!: 200,000 usage examples
Téarma.ie - National Terminology Database for Irish *www.téarma.ie*	An Coiste Téarmaíochta (Fiontar & Scoil na Gaeilge, DCU)	• Over 183,000 Irish terms
IATE Terminology Database *www.gaois.ie/trm*	Lex EU Project (Department of Culture Heritage and the Gaeltacht, Fiontar, DCU)	• Over 130,916 Irish legal terms
Logainm.ie *www.logainm.ie*	Placenames Database (Fiontar, DCU and Department of Culture, Heritage and the Gaeltacht)	• Idiom collection: 420 multiword expressions • 200,000 Irish-English placename pairs • 100,000 geo-tagged placenames
www.Ainm.ie	A database based on the series Beathaisnéis, featuring people who have had a substantial impact on the Irish Language.	• 1,769 curated biographies in linked data format

Table 4.1: An overview of the available linguistic data resources for Irish

[1] https://en.wikipedia.org/wiki/Knowledge_graph
[2] Over 100,000,000 searches to date (https://www.forasnagaeilge.ie/cead-milliun-cuardach-agus-cead-milliun-buiochas/)

Syntactic and semantic resources

In this section, we give an overview of the syntactic and semantic resources under development for Irish. These resources are mostly in the early stages of development and require substantial financial support to bring them to the level required to be effective in language applications.

Foclóir Briathra Gaeilge is a verb valency dictionary for Irish[1]. It provides structured data for approximately 200 Irish verbs, giving types, meanings and usage patterns and typical collocates (words that are used along with them). Figure 5 shows the valency information for the verb *abair* 'say'. This valency dictionary was developed in Germany by Prof Arndt Wigger, using Caint Chonamara, a large spoken corpus collected in 1964 [11] ,and Corpas Náisiúnta na Gaeilge [12] an eight million-word corpus of Irish written language. The data amounting to approximately 5,500 entries is available in an XML format and is available on the Potafocal.ie website (funded by An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta).

Computational grammars serve as the basis for syntactic and computational descriptions of the Irish language. They provide scientific descriptions of a language in which linguistic patterns and rules are expressed in a form that is computationally interpretable, so that language processing applications can be designed (see Chapter 5).

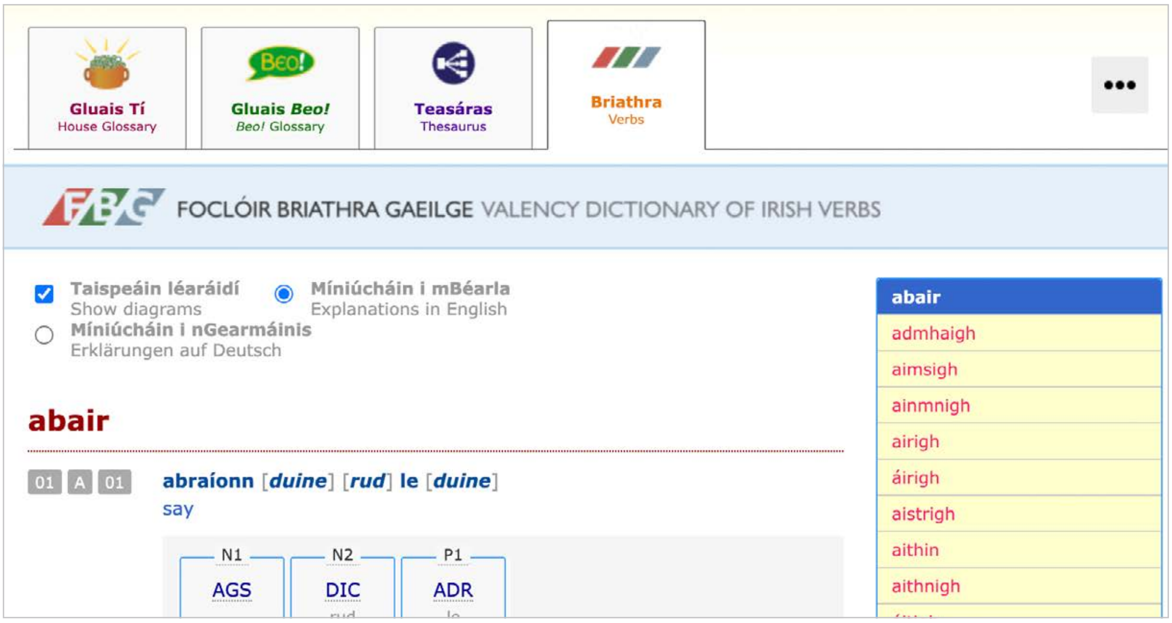
Some recent computational linguistic research in this area has documented many of the structures and patterns that occur in Irish [13-17], however, a comprehensive computational grammar for Irish still needs to be developed.

Líonra Séimeantach na Gaeilge (LSG) is a large-scale semantic network for Irish [18] that was created by mapping over word senses from the Princeton (English) WordNet into Irish, semi-automatically (see Figure 4.6). This means that it is automatically 'linked' to the English WordNet, and therefore, in turn, linked with semantic networks for dozens of other languages. A recent version of the LSG making these links explicit was produced as part of research on English-Irish machine translation [19].

As the LSG was based on the English WordNet, it will contain semantic mappings which are not found in Irish, and conversely may be missing mappings which are relevant to Irish. Further manual verification of the automatically generated semantic links between English and Irish is necessary, to ensure that the wordnet is tailored to Irish. Ongoing support is also necessary as new terms must have their semantic relationship to existing entries identified and labelled. Such a verified resource can be used as a thesaurus for writing in Irish (Chapter 13), and in applications such as Irish language information retrieval (Chapter 12). It could also form an essential component of a variety of NLP tools such as broad-coverage word sense disambiguation systems or semantic parsers (Chapter 5). The LSG[2] database is open source and freely available for use by researchers and NLP software developers.

Polylingual Wordnet[1]

Polylingual Wordnet uses existing Wordnets from



☑ Taispeáin léaráidí
Show diagrams

● Míniúcháin i mBéarla
Explanations in English

○ Míniúcháin i nGearmáinis
Erklärungen auf Deutsch

abair

01 A 01 abraíonn [*duine*] [*rud*] le [*duine*]
say

Figure 4.5: Foclóir Briathra Gaeilge (Irish Verb Valency Dictionary)

[1] http://www.potafocal.com/fbg/
[2] https://cadhan.com/lsg/

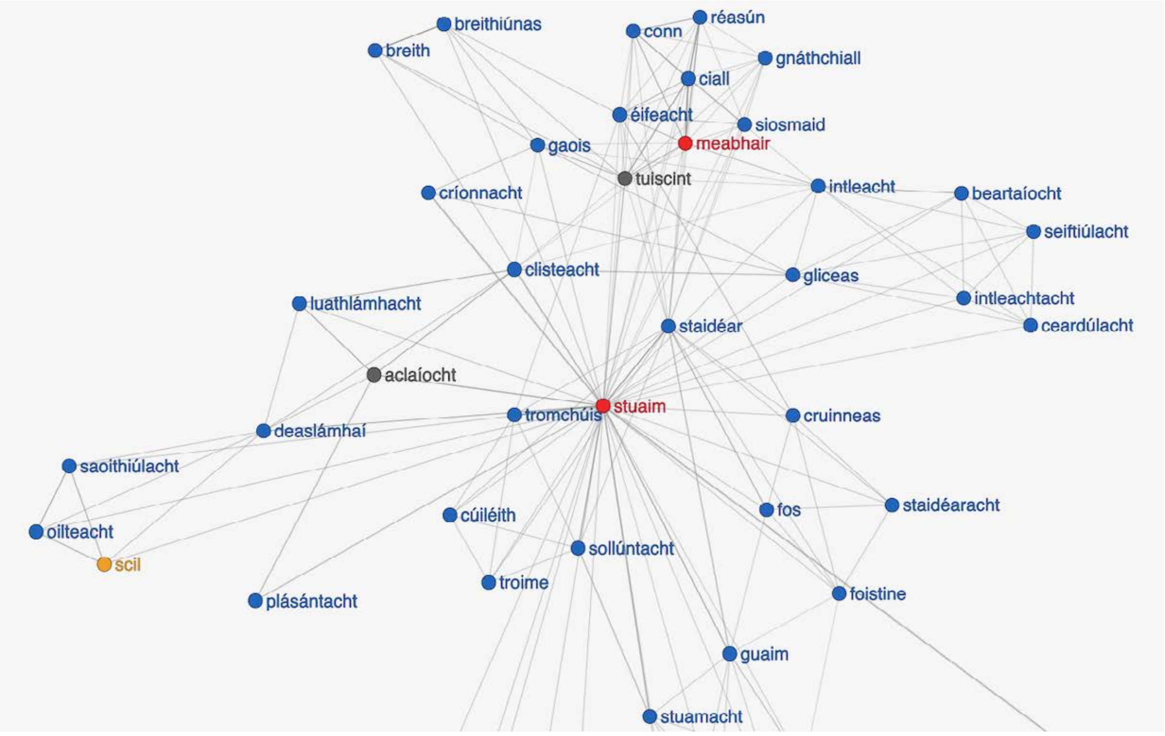


Figure 4.6 : LSG (Irish Wordnet) representation for the term [stuaim] *'steadiness'*

Open Multilingual Wordnet[2] and DBpedia[3] to generate new wordnets for additional languages. This approach has been used to generate wordnets for more than 20 languages including Irish [20]. This would also require manual verification of automatically generated links to ensure its accuracy for downstream applications.

DBpedia as Gaeilge is the chapter responsible for the maintenance of the Irish language version of DBpedia. This chapter extracts data from An Vicipéid[4] (Irish Wikipedia), making it available to query and link to other datasets. It also translates the English terms from the DBpedia ontology. This means that any application created from such ontologies can be easily localised to Irish in the future.

Data from Logainm.ie (see Table 4.1), which is available as Open Linked Data, is now in use by Ordinance Survey Ireland, Eircode, GeoDirectory, OpenStreetMap (Figure 8), and many others to allow for a localised Irish language query. A resource like this is important for NLP tasks such as Named Entity Recognition (see Chapter 5) and any intelligent application that requires geo-location identification (Chapter 13).

4.5 Language Resources: Knowledge Bases – General Recommendations

Knowledge bases are fundamental resources in the development of language tools and applications. While some progress has been made in the past with such resources, we provide a summary here of the requirements for ensuring the Irish language has a strong basis on which to further develop the necessary tools and applications identified in this document – firstly with a general recommendation, followed by recommendations per knowledge base type:

- User interfaces require ongoing hosting and maintenance to ensure maximum return on the investment in creating these valuable resources, and continued availability to the public.
- Database infrastructures require regular maintenance and updating in order to keep up with state-of-the-art and changing technology.
- Development of new knowledge bases should be application-led to ensure optimum relevance and usability.
- Lexical databases and domain-specific terminology databases need to be kept up to date with newly identified and documented

[1] http://polylingwn.linguistic-lod.org/
[2] http://compling.hss.ntu.edu.sg/omw/
[3] http://ga.dbpedia.org/
[4] https://ga.wikipedia.org/wiki/Pr%C3%ADomhleathanach

words and terminology in the Irish language.

- As the number of independent lexical knowledge bases for Irish continues to grow, greater coordination and data-sharing will be required to ensure consistency and compatibility. For example, the spelling of a small number words varies between focloir.ie, téarma.ie, and the Foclóir Gaeilge-Béarla, which can lead to confusion amongst users and systems alike. At a minimum, maintainers of these databases should coordinate when orthographic or grammatical changes are implemented; ideally, all of the databases should be linked together to allow automated consistency checks.
- An integrated search facility for public use which searches across multiple databases e.g. téarma.ie, focloir.ie and teanglann.ie. needs to be developed.
- Crowdsourcing should be encouraged where appropriate, as in the case of Irish Placenames data which can be submitted by the public to MeithealLogainm.ie (see Chapter 15).

Lexical Resources – Required Future Developments

Some of the more urgent requirements include:

- New domain-specific terminology databases are required within computer-assisted translation environments, e.g. in the legal field, science teaching material.
- An up-to-date monolingual Irish dictionary is necessary for use in Irish medium schools, and Irish medium workplaces.
- An interactive database of Irish multi-word expressions (e.g. idioms, light verb constructions, particle verbs, noun compounds) that document the usage of words in the context of frequently occurring surrounding words, e.g. Ilfhocal [21].
- Further development of Logainm.ie database, primarily to add townland names and street names, and its web service functionality to be enhanced.
- Use of terminometry to determine usage (or non usage) of terms over time to ensure databases are up-to-date (e.g. féinphic vs. féinín for 'selfie')
- Ongoing maintenance of dictionaries and termbases through the extraction of terms from corpora (see Chapter 3).
- Specifically tailored tools and methods to help with terminology extraction from corpora and translation memories.

Syntactic and semantic Resources – Required Future Developments

Some of the more urgent requirements include:

- Development of computational grammars using established computational theoretical syntactic frameworks such as Lexical Functional Grammar, Head-driven Phrase Structure Grammar, Combinatory Categorial Grammar, or similar, to underpin parsing technology.
- A more comprehensive valency dictionary, building on Foclóir Briathra Gaeilge (FBG) where possible, and using more recent corpus data.
- Development of an Irish VerbNet-type resource in conjunction with the development of lexical resources such as FBG and LSG (see Semantic Resources).
- Development of a FrameNet-type resource that would complement a valency dictionary (see above), in identifying verb usage patterns, to be used in rule-based syntactic/semantic parsing and rule-based elements of machine translation
- Ongoing support and updates for all semantic networks such as Líonra Séimeantach na Gaeilge (LSG) or Polylingual Wordnet (Irish), including manual verification of the automatically generated mappings in semantic networks in order to enhance the accuracy to ensure reliability in end-user applications such as Thesauri.
- Further research into automatic induction of semantic networks which would improve existing links, and make the addition of new entries easier.
- More linked data repositories of Irish content such as DBpedia as Gaeilge.
- There is a need for the creation of ontologies and taxonomies that can support the development of knowledge graphs.
- No research has been carried out for Irish in the growing area of knowledge bases. Fundamental research is required to avail of the opportunities these tools can provide.

References

[1] Ó Dónaill, N. (1977). *Foclóir Gaeilge-Béarla*. Dublin: Oifig an tSoláthair.

[2] Ó Dónaill, N. & Ua Maoileoin, P. (1991). *An Foclóir Beag*. Dublin: Oifig an tSoláthair.

[3] de Bhaldraithe, T. (1959). *Foclóir Béarla-Gaeilge*. Dublin: Oifig an tSoláthair.

[4] Maldonado, A. & Lewis, D. (2016). Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In *Proceedings of the 12th International Conference on Terminology and Knowledge Engingeering* (pp. 91-100).

[5] Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006). A large-scale extension of VerbNet with novel verb classes. In *Proceedings of the 12th EURALEX International Congress* (pp. 173-184).

[6] Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006). Extensive classifications of English verbs. In *Proceedings of the 12th EURALEX International*

Congress (Vol. 4).
[7] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

[8] Baker, C.F., Fillmore, C.J. & Lowe, J.B. (1998). The Berkeley FrameNetProject. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (Vol.1, pp. 86-90).

[9] Petruck, M.R.L. & Kordoni, V. (2015). Robust Semantic Analysis of Multiword Expressions with FrameNet. In *Conference on Empirical Methods in Natural Language Processing 2015*.

[10] Johannsen, A., Martínez Alonso, H., & Søgaard, A. (2015). Any-language frame-semantic parsing. In *Conference on Empirical Methods in Natural Language Processing 2015* (pp. 2062-2066).

[11] Wigger, A. (2000). *Caint Chonamara: Bailiúchán Hans Hartmann*. University of Bonn.

[12] Institiúid Teangeolaíochta Éireann. (2000). *Corpus Náisiúnta na Gaeilge*. Dublin.

[13] Uí Dhonnchadha, E. (2009). *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. [Ph.D. thesis]. Dublin City University.

[14] Lynn, T. (2016). *Irish Dependency Treebanking and Parsing*. [Ph.D. thesis]. Dublin City University & Macquarie University.

[15] Sulger, S. (2009). *Irish Clefting – The LFG Perspective*. [Ph.D. thesis]. University of Konstanz.
[16] Cassidy, L., Lynn, T., Barry, J. & Foster, J. (2022). TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 6869-6884).

[17] Lynn, T. & Foster, J. (2016). Universal Dependencies for Irish. In *Proceedings of the 2nd Celtic Language Technology Workshop* (pp. 79-92).

[18] Scannell, K. (2007). *Líonra Seimeantach na Gaeilge*.

[19] O'Regan, J., Scannell, K. & Uí Dhonnchadha, E. (2016). lemonGAWN: WordNet Gaeilge as Linked Building and Using Linked Language Resources. In *Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources* (pp. 36-41).

[20] Arcan, M., McCrae, J. & Buitelaar, P. (2019). *Polylingual WordNet*. https://arxiv.org/list/cs/recent.

[21] Walsh, A., Lynn, T. & Foster, J. (2019). Ilfhocail: A Lexicon of Irish MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet* (pp. 162-168).

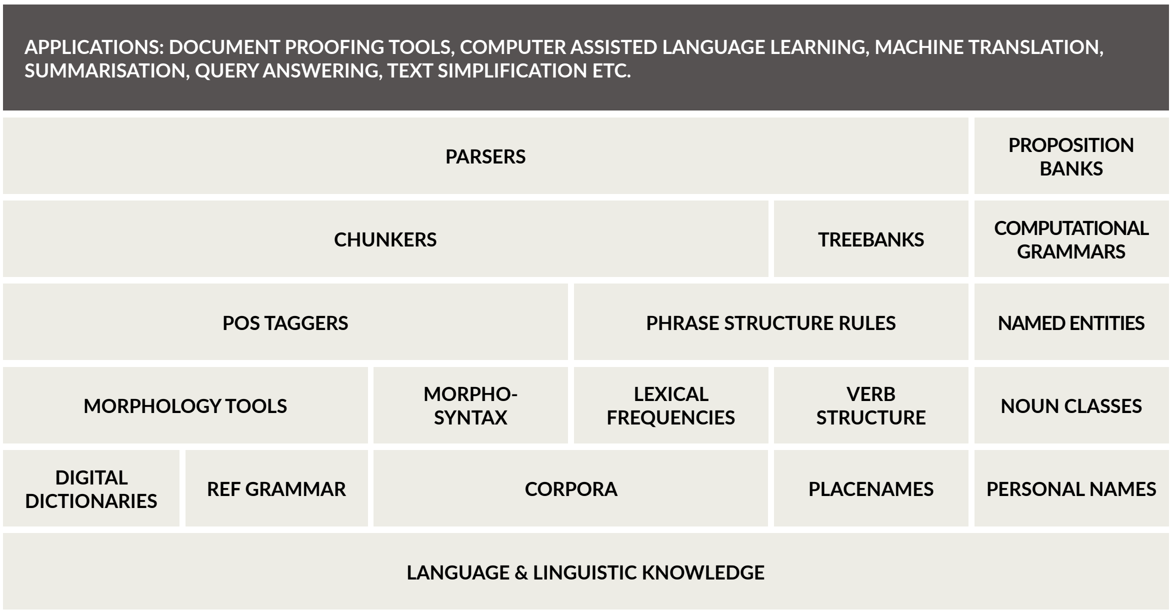
Chapter 5
Natural Language Processing Tools

5.1 What is Natural Language Processing?

Natural Language Processing (NLP) is the term used for the automatic processing of human language by a computer. Processing includes trying to interpret the meaning of written or spoken language, and the generation of human language in answer to a query, as part of a dialogue, or as a translation. Human language is a highly complex phenomenon; therefore, the automatic processing of language is usually broken down into many small steps. Each of these steps is tackled using software which addresses a specific task. The pieces of software that carry out these tasks are commonly known as Natural Language Processing tools. These tools, together with language resources such as corpora (Chapter 3) and knowledge bases (Chapter 4), are the essential building blocks of large complex systems detailed in this plan. Figure 5.1 illustrates some of the key resources and tools which are used in Natural Language Processing.

5.2 Why is this important and for whom?

In order to carry out every day word-processing tasks such as spell-checking and grammar checking we need to be able to analyse the words, phrases and sentences in a text, and we need to be able to make lexically and grammatically correct suggestions. In CALL systems for education, we also need to be able to assess the language input and generate appropriate responses. In order to provide a user-friendly means of human-computer communication we will require a wide range of NLP tools for morphology analysis and generation, part-of-speech (POS) tagging, chunking, syntactic and semantic parsing, etc. which are described in more detail in the next section. One of the first NLP applications for modern Irish was the annotation of a corpus[1] of Irish texts with part-of-speech tags to facilitate the development of the New English-Irish Dictionary[2].



APPLICATIONS: DOCUMENT PROOFING TOOLS, COMPUTER ASSISTED LANGUAGE LEARNING, MACHINE TRANSLATION, SUMMARISATION, QUERY ANSWERING, TEXT SIMPLIFICATION ETC.

PARSERS | PROPOSITION BANKS

CHUNKERS | TREEBANKS | COMPUTATIONAL GRAMMARS

POS TAGGERS | PHRASE STRUCTURE RULES | NAMED ENTITIES

MORPHOLOGY TOOLS | MORPHO-SYNTAX | LEXICAL FREQUENCIES | VERB STRUCTURE | NOUN CLASSES

DIGITAL DICTIONARIES | REF GRAMMAR | CORPORA | PLACENAMES | PERSONAL NAMES

LANGUAGE & LINGUISTIC KNOWLEDGE

Figure 5.1: NLP Resources and Tools

[1] http://corpas.focloir.ie/
[2] https://focloir.ie/

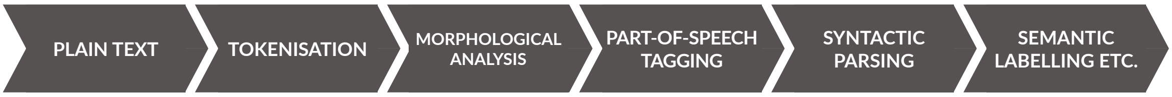


Figure 5.2: NLP annotation of text for language applications

5.3 How does it work?

In order to process a sentence, the typical NLP tasks could include dividing the sentence into individual words (and some multi-word units such as place names etc.), assigning part-of-speech categories to the words to determine the nouns and verbs etc., then assigning morphological and semantic features to the words, e.g. singular/plural for nouns, tense for verbs, etc. The words then need to be grouped into phrases, and the relationships between the phrases must be determined in order to discover the basic meaning, i.e. the semantic roles of the entities involved, which tells us 'who did what to whom'. This sequence of events, as illustrated in Figure 5.2, requires a number of NLP tools, which typically include a tokeniser, a morphological analyser, a part-of-speech tagger, parser, a semantic role labeller, etc. Each of these tools adds additional information to the initial plain text sentence in the form of 'annotations' or 'labels'.

This is largely a rule-based method of processing language. When sufficient resources are in place, statistical and machine-learning methods can also be used to great effect. For example, when sufficiently large bodies of annotated texts (or parallel corpora of aligned plain texts in the case of translation) have been accumulated and quality checked, machine-learning methods and tools can be used, whereby the system can begin to discern patterns and derive rules automatically. Recent advances in deep learning have resulted in the use of pre-trained word embeddings (mathematical representations of words), which provide insight into language use in context and therefore greatly improve performance.

5.4 What has been done to date for Irish?

Morphology tools

Morphology tools include tools for tokenisation lemmatisation, morphological analysis and generation. Tokenisation is the task of separating the sentence or utterance in to units called tokens which include words, punctuation and multi-word items, while lemmatisation is the task of identifying the dictionary headword associated with inflected forms, e.g. *cuir* 'put' is the lemma for *cuirfidh* 'will put' and *cuireann* 'puts' etc. A finite-state rule-based morphological analyser and generator for Irish [1] uses hand-encoded morphology rules based on grammar paradigms [2, 3] and has a large internal lexicon which includes the headwords and variants found in the Ó Dónaill dictionary [4]. These morphology tools identify the lemma, a part-of-speech category and associated features of each token in a text. Many words can have more than one possible part-of-speech and/or lemma, and morphological analyser outputs all possible analyses for each word (token). These tools also include a component that predicts the most likely part-of-speech category for unknown words based on their external characteristics, e.g. recognisable prefixes and suffixes, or broad/slender endings. For example, if an unknown word ends in *–faidh* the morphological analyser will predict that it is a verb in the future-tense, even though the root or lemma is not in the lexicon.

For rule-based tools such as these, it is important to keep the internal lexicon up to date. New words and terminology regularly come into the language (see Chapter 3 regarding corpora and and Chapter 4 regarding dictionaries and lexical knowledgebases), therefore the internal lexicon needs to be regularly updated with new words and terminology. The morphology lexicon can be enhanced through the addition of more multi-word items, e.g. *ar cosa in airde* 'at a gallop', and idioms such as *ar muin na muice* 'on the pig's back'.

ar	is	Copula
ar	ar	Noun+Masc+Com+Sg
ar	ar	Verb+PastInd
ar	ar	Prep+Simp
ar	ar	Particle+Verbal+Q+Past
ar	ar	Particle+Verbal+Rel
muin	muin	Noun+Fem+Com+Sg
na	na	Article+Pl+Def
na	na	Article+Gen+Sg+Def+Fem
muice	muc	Noun+Fem+Gen+Sg

Figure 5.3: Morphological analysis of 'ar muin na muice' showing the token in column 1, the lemma in column 2, and part of speech category + features in column 3

Lynn et al's [5] study on the Irish language use on Twitter involved the application of a statistical or data-driven lemmatiser, Morfette [6]. By training the tool on manually verified lemmas (roots) of tweet contents, it learned to predict the lemma of over 97% of words in previously unseen tweets. Further experiments are required using statistical analysis with data-driven morphological tools such as Morfette or MARMOT [7] on a large dataset of both standard Irish text and noisy user-generated text, e.g. social media data.

Part-of-Speech Tagging Tools

The purpose of a part-of-speech (POS) tagger is to assign the appropriate POS tag and features to each word in a sentence. The challenge for POS tagging is that most words can function in a variety of ways , e.g. *glan* 'clean' might function as an adjective in *seomra glan* 'a clean room' or it might function as a verb in *ghlan sé é* 'he/it cleaned it'. A rule-based POS tagger [8] takes the choice of possible analyses provided by the morphological analyser (see above), and applies hand-coded syntactic rules in order to determine the appropriate POS category for a word in a particular context, giving an outcome as in Figure 4.

ar	ar	Prep+Simp
muin	muin	Noun+Fem+Com+Sg
na	na	Article+Gen+Sg+Def+Fem
muice	muc	Noun+Fem+Gen+Sg

Figure 5.4: Lemmatisation and POS tagging for 'ar muin na muice'

In cases where the rule-based system cannot make a decision based on local syntactic context alone, lexical frequency information can be used to determine the most likely POS category. For example in a particular sentence, where the word *siad* might be a pronoun meaning 'they' or it might be a noun meaning 'growth or swelling', the more frequent pronoun interpretation of *siad*, 'they' would be favoured based on lexical frequency information derived from a corpus of Irish (Chapter 3). Hybrid approaches of rule and statistical methods have already proven successful in the development of tools for other languages [9]. In addition, when sufficient manually checked POS-tagged data becomes available, a statistical POS tagger for Irish could be trained on this gold-standard POS-tagged data (see Chapter 3).

Some domain-specific experiments in statistical POS tagging have been carried out on Twitter data using a statistical analyser trained on a small Tweet corpus [5]. A benefit of this domain-specific training is that it enables easier processing of non-standard language, e.g. Twitter data.

Chunkers and Phrase-based Tools

The goal of a chunker is to group words into phrases by identifying the head word and the words which are dependent on the head word, e.g. in a noun phrase, there will be a head noun and there may be variety of words which are dependents of the noun, such as an article, or adjectives or a modifying nouns. Phrases or chunks, rather than individual words, each have a grammatical role in the sentence, i.e. a noun-phrase will function as the grammatical 'subject', or 'direct object' etc.. A rule-based dependency chunker (partial parser) for Irish [8] identifies phrases and their grammatical functions (e.g. subject, direct object etc.). To extend this to full rule-based syntactic dependency parser requires further research on verb-argument structure [10] and the semantic properties of nouns, (see Chapters 3 and 4).

```
Bhí na buaiteoirí ar muin na muice.

[S   [V Bhí bí+Verb+VI+PastInd+Len+@FMV]
     [NP na na+Art+Pl+Def+@>N
          buaiteoirí buaiteoir+Noun+Masc+Com+Pl+@SUBJ]
     [PP ar ar+Prep+Simp+@PP_SC
          [NP muin muin+Noun+Fem+Com+Sg+@P<
               na na+Art+Gen+Sg+Def+Fem+@>N
               muice muc+Noun+Fem+Gen+Sg+@N] ] ]
```

Figure 5.5: Sentence with bracketed phrases (or chunks) and grammatical functions and dependencies

Syntactic Parsers

A parser gives a full syntactic analysis of a sentence. It identifies all of the relationships between phrases and the role that they play in describing an event. Full syntactic statistical parsers have been developed for Irish [11], which connect all words in a sentence through meaningful grammatical roles known as dependency relations, see Figure 5.6. These syntactic parsers are data-driven and rely on data from the Irish Dependency Treebank and the Irish Universal Dependency Treebank (see Chapter 3) to induce and learn grammatical structures in the language. The parsers then, based on probabilities of a combination of features in the treebank (word forms, lemmas, morphological features, previously seen dependencies) can predict the syntactic structure of previously unseen sentences. As with the nature of data-driven tools, the parser's accuracy will increase with the growth of the treebanks.

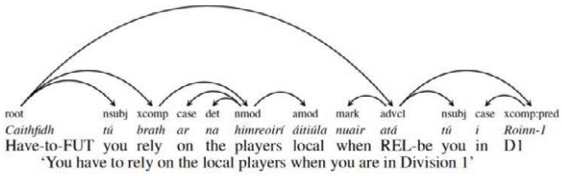

root
nsubj
xcomp
case
det
nmod
amod
mark
advcl
nsubj
case
xcomp:pred
Caithfidh tú brath ar na himreoirí áitiúla nuair atá tú i Roinn-1
Have-to-FUT you rely on the players local when REL-be you in D1
'You have to rely on the local players when you are in Division 1'

Figure 5.6: Sentence with syntactic dependency tree showing grammatical functions and dependencies

Research on parsing approaches other than dependency parsing (e.g. constituency parsing, Combinatory Categorial Grammar (CCG) parsing) may also prove useful for some downstream tasks. Some preliminary work on CCG for Scottish Gaelic [12] could help efforts for Irish treebank research in this regard.

5.5 Future Directions

In addition to ongoing development and maintenance of tools as described in the previous sections, the following sections give an overview of some of the NLP tools that are available for other languages, but have not yet been developed for Irish. We summarise their importance here and their relevance in end-user applications. As this is a rapidly evolving field, we can expect new tools and methodologies to emerge in the coming years. It is therefore essential that Irish stays abreast of international developments.

Semantic Role-labelling and Semantic Parsing

A parser assigns a syntactic structure and grammatical roles to a sentence, but for many tasks it is necessary to know the semantic roles associated with the entities in the sentence in order to be able to know 'who did what to whom'. Many natural language understanding applications such as information extraction and summarisation, question answering and sentiment analysis etc. must have access to this type of semantic information.

(1)	Thug	Seán	leabhar	dá	chara
	Gave	Seán	book	to-his	friend
		AGENT	THEME		RECIPIENT
	REL	ARG0	ARG1		ARG2

'Seán gave a book to his friend'

Figure 5.7: Sentence with semantic role labels.

A semantic role labeller (SRL) usually assigns semantic roles to the verb arguments and modifiers. For example, in Figure 5.7 Seán has the 'agent' role, i.e. the 'doer' of the action, *chara* 'friend' has the 'recipient' role and *leabhar* 'book' has the 'theme' role. Once a computer can decipher these components of a sentence [13, 14], it is in a better position to carry out tasks such as translation, summarisation, error detection, and so on.

The work carried out on verb valency in Foclóir Briathra Gaeilge (FBG) (see Chapter 4) could provide a basis for the development of a semantic role-labelling tool.

Word sense disambiguation

Word sense disambiguation (WSD) is an essential task in many applications involving information extraction, machine translation etc. Many words have more than one meaning and the challenge is to choose the correct meaning in a particular context. For example, the word *caith* has a number of different meanings, e.g. it could mean 'throw', 'spend', 'wear', or 'smoke'. In NLP applications, such as information extraction, document summarisation and indexing, and machine translation etc., it is essential that the appropriate sense is identified in order to provide accurate responses to a query.

There has been a significant increase in the use of WordNets [15] as well as the use of neural networks and word embeddings [16] in NLP tools that exploit machine-learning approaches. From this perspective, the semantic resources including Irish Líonra Séimeantach na Gaeilge (LSG) (see Chapter 4) could provide a basis from which to develop the state-of-the art WSD tools currently in use for other languages.

Named entity recognition

Many of the unknown words in a text tend to be proper nouns referring to people, places and organisations. A named-entity recogniser (NER) identifies and classifies words and phrases as names of organisations, people, places, elements of time and so on. In example (2), the entity types in this sentence have been labelled as an 'organisation' and as a 'date'.

(2)	Bunaíodh	[Grúpa	Gnímh	an	N59]	san oíche	[Dé Céadaoin]
		ORG					DATE

established [group action-GEN the N59] in-the night [Wednesday]
'The N59 Action Group was established Wednesday night'

Figure 5.8: Sentence with entity annotations.

An NER tool is extremely valuable for tasks such as text summarisation, information retrieval and translation [17, 18]. To date, there is no named-entity recognition system available for Irish. There are some basic resources (lists of named entities) available through the part-of-speech tagger technology, and place names at logainm.ie but much work is required to extend this research into a comprehensive NER tool.

Sentiment analysis

Sentiment refers to the intended meaning or emotion conveyed in text or speech, for example, sentiments such disgust, anger, surprise and joy. Often sentiment analysis tools focus on whether a positive, negative or neutral sentiment is being conveyed. This technology can be applied in many contexts, such as a government gauging the level of support from the public on a certain topic, or a business assessing the market's opinion on their products or services. Sentiment analysis is often carried out nowadays on online content for these purposes [19]. However, sentiment analysis can also assist in the area of digital humanities, where historical researchers search for articles or documents that negatively or positively report on or portray a particular subject.

A sentiment analysis project carried out by researchers in DCU reported on public opinion in the lead up to the 2016 General Election, through the analysis of Irish tweets [20].

Text simplification

The purpose of a text simplification tool is to analyse a text with a view to making it simpler and more understandable to the reader, where possible. Text simplification techniques focus on simplifying both the lexicon and/or the syntactic structure. Text simplification has many applications in the area of education [21, 22, 23] and public administration [24]. These types of tools are particularly relevant in the context of providing texts for learners of Irish in the education system, and the provision of documents for members of the public in their dealings with local and central government.

5.6 Recommendations

- Existing tools require regular maintenance and updating in order to keep up with state-of-the-art and changing technology, including software updates.
- The development of new tools should be both research-led and application-led to ensure optimum relevance and usability.
- Funding should be available for ongoing hosting and maintenance of user interfaces to ensure maximum access to these valuable resources.
- The establishment of a central repository of lexical and terminological data in a common format (see Chapter 4), is necessary to enable developers to keep existing NLP tools to up-to-date in a routine and flexible manner.
- Many of the existing and new NLP tools described in this chapter, depend on the theoretical underpinnings in Chapter 2 together with the availability of sufficiently large and accurate lexical knowledge bases and corpora as described in Chapters 3 and 4.
- The improvement of part-of-speech tagging and parsing of both published and user-generated content such as Facebook, Twitter etc. requires larger gold-standard corpora (see Chapter 3)
- The improvement of rule-based dependency parsing for Irish relies on the development of lexical resources such as a verb valency dictionary and noun semantic class information (see Chapter 4).
- The improvement of statistical dependency parsing for Irish relies on extending the Irish Treebanks (see Chapter 3) along with enhanced parsing techniques.
- The development of new NLP tools for semantic role-labelling (SRL) and word sense disambiguation (WSD), requires lexical resources and annotated corpora (see Chapters 3 and 4).
- The development of named entity recognition (NER) tools requires chunkers/syntactic parsers, and lexical resources including logainm.ie for Irish place names.
- The development of sentiment analysis tools requires the extension of the current lexicon of polarity-labelled words [20], along with a broadening of this lexicon to other domains.
- The development of text simplification tools requires syntactic theory (Chapter 2) together with lexical and syntactic resources (Chapter 4 and this chapter).

References

[1] Uí Dhonnchadha, E. (2002). *An Analyser and Generator for Irish Inflectional Morphology using Finite-State-Transducers*. [M.Sc. thesis]. Dublin City University.

[2] Congregation of Christian Brothers. (1999). *Graiméar Gaeilge na mBráithre Críostaí* (2 ed.). Dublin: An Gúm.

[3] Congregation of Christian Brothers. (1988). *New Irish Grammar*. Dublin: C.J. Fallon.

[4] Ó Dónaill, N. (1977). *Foclóir Gaeilge Béarla*. Dublin: Oifig an tSoláthair.

[5] Lynn, T., Scannell, K. & Maguire, E. (2015), Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the 1st Workshop on Noisy User-generated Text*.

[6] Chrupała, G., Dinu, G. & van Genabith J. (2008). Learning Morphology with Morfette. In *Proceedings of LREC 2008*.

[7] Müller, T., Schmid, H., & Schütze, H. (2013). Efficient higher-order crfs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

[8] Uí Dhonnchadha, E. (2009). *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. [Ph.D. thesis]. Dublin City University.

[9] Loftsson, H., Helgadóttir, S., and Rögnvaldsson, E. (2011). Using a morphological database to in-crease the accuracy in pos tagging. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 49–55).

[10] Wigger, A. (2008). Advances in the lexicography of Modern Irish verbs. In A. Bloch-Rozmej (Ed.), *Issues in Celtic Linguistics* (pp. 233-250).

[11] Lynn, T. (2016). *Irish Dependency Treebanking and Parsing*. [Ph.D thesis]. Dublin City University & Macquarie University.

[12] Batchelor, C. (2014). gdbank: The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic. In *Proceedings of CLTW2014, 1st Celtic Language Technology, Workshop*

XVII.

[13] Gildea, D., & Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics 28*(3), 245-288.
[14] Pradhan, S.S., Ward, W., & Martin, J. H. (2008). Towards Robust Semantic Role Labeling. *Computational Linguistics 34*(2), 289-310.

[15] Miller, G.A. (1995). WordNet: A Lexical Database for English. *ACM 38*(11), 39-41.

[16] Saedi, C., Branco, A., António Rodrigues, J. & Silva, J. (2018). WordNet Embeddings. [Paper presentation]. *Proceedings of The Third Workshop on Representation Learning for NLP*.

[17] Toral, A., Ferrández, S., Monachini M. & Muñoz, R. (2012). Web 2.0, Language Resources and standards to automatically build a multilingual Named Entity Lexicon. *Language Resources and Evaluation 46*(3), 383-419.

[18] Eiselt, A & Figueroa, A. (2013). A Two-Step Named Entity Recognizer for Open-Domain Search Queries. In *Proceedings of the International Joint Conference on Natural Language Processing* (pp. 829–833).

[19] Mäntylä, M., Graziotin, D. & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review* 27, 16-32.

[20] Afli, H., McGuire, S., & Way, A. (2017). Sentiment translation for low resourced languages: Experiments on Irish general election tweets. [Paper presentation]. *18th International Conference on Computational Linguistics and Intelligent Text Processing*.

[21] Petersen, S.& Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. In *Proceedings of SLaTE-2007*, 69-72.

[22] Bott, S., Saggion, H. & Mille, S. (2012). Text Simplification Tools for Spanish. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

[23] Leal, S., Duran, M. & Aluísio, S. (2018). A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics.*

[24] Scarton, C., Paetzold, G. & Specia, L. (2018). Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Chapter 6
Natural Language Generation

6.1 What is NLG?

Natural Language Generation (NLG) is the branch of Natural Language Processing that concerns generating natural language from non-linguistic data sources, such as numerical data, semantics such as some logical form or structured knowledge bases [1, 2]. While NLG may include text-to-text systems i.e. abstractive text summarisation or the generation of textual descriptions from image or video [2], the focus here is on the type of NLG known as Data-to-Text systems [1]. Quite simply, we can view an NLG system as a translator that converts data (e.g. numerical data from a weather report or financial portfolio) into a tailored natural-language summary (e.g.) a weather or personalised financial report.

A simple example is a system that generates letters. These do not typically involve grammar rules, but may generate a letter to a consumer, e.g. stating that a credit card spending limit was reached. To put it another way, simple (or shallow) NLG systems use a template, not unlike a Word document mail merge, but more modern complex NLG systems dynamically create the text. As in other areas of language technology, this can be done using either explicit models of language (e.g. grammars and rule sets), or using statistical models derived by analysing existing data, ie. human-written text.

In parsing, the system needs to disambiguate the input sentence to produce the more structured machine or human readable representation of natural language. In NLG the system instead needs to make decisions about how to put a concept into words.

Natural Language Generation (NLG) vs. Understanding (NLU)

NLG has existed for a long time but commercial NLG technology has really only recently become widely available. NLG has followed a similar historical path to NLU and machine translation (Chapter 11), albeit quite delayed, with the advent of evaluation efforts across the research community, data-driven approaches and the eventual offering of commercial NLG systems. This delay is the result of the challenging nature of the NLG problem. In NLU, the task is better defined and its input – natural language – is known and widely available on the Web. However, with respect to NLG, the output (text) is known, while historically there had been a deficit of appropriate input (data) suitable for generation [3]. This, in conjunction with the historical absence of suitable data-to-text corpora, made it challenging to agree on evaluation approaches and NLG frameworks. The rise of Big Data coupled with the recent upsurge in deep neural network approaches to NLP, have caused renewed interested in language generation, with the emergence of data-to-text datasets [4, 5], which use variations of Abstract Meaning Representation (AMR) [6] as an agreed input format. Despite this progress the majority of available open source NLG components tend to be scattered, obsolete or propriety. Therefore, one promising research avenue is the development of an open common platform or suite of reusable tools for NLG which could help close the research gap between NLU and NLG.

6.2 Why is it important and for whom?

In our current Big Data era, the global data sphere continues to grow exponentially, with an estimated tenfold increase to 163 zettabytes (zb, that is a trillion gigabytes) by 2025[1]. Big Data Analytics is hurtling towards an inevitable knowledge-access bottleneck, whereby data analytics will have analysed and aggregated data, generating lucrative but frustratingly inaccessible insights for the non-expert. Natural Language Generation (NLG) has been described as one of the last miles in the Big Data race[2].

From a commercial perspective, the most successful NLG applications have been data-to-text systems which generate textual summaries of databases and data sets; these systems usually perform data analysis as well as text generation. In particular, several

[1] DELIVERING INFORMATION. FROM BIG DATA. An NLG Thought Leadership White Paper by Dr. Robert Dale, NLG Chief Strategy Scientist. https://www.arria.com/wp-content/uploads/2017/09/White-Paper-Big-Data-ARR0035.pdf

[2] Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big. An IDC White Paper, Sponsored by David Reinsel John Gantz John Rydning | April 2017

systems have been built that produce textual weather forecasts from raw weather station data. The earliest such system to be deployed was FoG [7], which was used by Environment Canada to generate weather forecasts in French and English in the early 1990s. The success of FoG triggered other work, both research and commercial. Recent research in this area includes an experiment which showed that users sometimes preferred computer-generated weather forecasts to human-written ones, in part because the computer forecasts used more consistent terminology [8,9] and a demonstration that statistical techniques could be used to generate high-quality weather forecasts. Recent applications include the UK Met Office's text-enhanced forecast [10].

The field of text summarisation relies heavily on NLG. This type of application is seen in context such as document summarisation (summarising a news article or report) as well as the snippets that are created in information retrieval applications (see Chapter 12), whereby search results are provided to a user and a short summary of the document found (e.g. a web page) is provided so that the user can decided whether it is worth clicking on the link provided.

In an Irish context, NLG technology can fill a gap by creating Irish language content on the fly for consumption through digital media. We've seen above how this technology can be used, very effectively, to provide weather forecasts from raw data. The same techniques can be applied by Met Éireann and other service providers whose services lend themselves to data e.g. traffic reports. NLG techniques have also been used very effectively to automatically produce reports on financial and other structured data sources. Following this model there is a wealth of opportunity to rapidly produce Irish language content that can respond in real time to changes in the underlying information. This way, public service announcements, reports and emergency information can be immediately made available to both language communities in Ireland without the need for translation or localisation. Text generation is also a valuable component in the development of chatbots.

6.3 How does it work?

The process of text generation can in its simplest form be as simple as reusing stock text variables which is linked together with some static 'glue text' to form new content in a process not dissimilar to mail merge (e.g. Dear [NAME], We received your registration request on [DD/MM/YYYY]). The results of this approach can be satisfactory for some applications but are far from sophisticated, scalable or portable between domains and tasks. A fully fledged NLG system typically includes various stages of intelligently designing such information merging to enable the generation of fluent, natural-looking text.

NLG Tasks and Architectures

Data-to-text NLG systems can be broken down into a number of tasks, which are often found in most NLG systems. These include:

- Content Determination: Deciding what information from the database should be included in the text being generated; i.e. what is salient. For instance, in a weather report, heavy rainfall data might indicate that a flood warning should be generated.
- Text Structuring (also referred to as Document or Text Planning): Determining the order in which information (called messages/events) will be presented in the text; Grouping of sentences in a generated text (for example into paragraphs).
- Sentence Aggregation: This involves deciding what information should be presented in individual sentences; i.e. deciding what messages/events should be merged into one sentence. The goal is to merge similar information to prevent repetition and improve readability and text fluidity. This involves merging syntactic constituents (see Chapter 5) together such as sentences and phrases.
- Lexicalisation: Choosing the right words and phrases to express information; This involves choosing the content words (nouns, verbs, adjectives, and adverbs) in a generated text. The

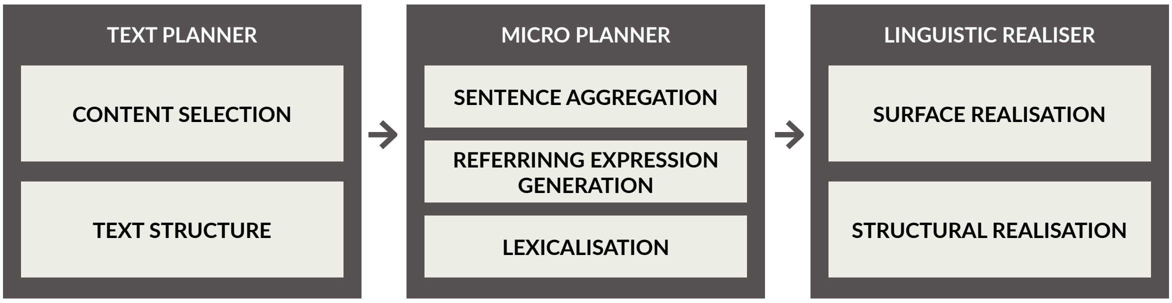


Figure 6.1: NLG classic three stage architecture

simplest type of word choice involves mapping a domain concept (perhaps represented in an ontology) to a word. A more complex situation is when a specific concept is expressed using different words or paraphrases in different situations (e.g. mathematical threshold vs door threshold).
- Referring expression generation: Selecting the words and phrases to refer to objects or entities (e.g. These results, that statement, them, it, etc.)
- Linguistic Realisation converts abstract representations (words and phrases) into well-formed sentences, according to the rules of syntax, morphology, and orthography (Chapter 2). This step combines Surface Realisation with the use of appropriate mark-up (Structural Realisation).

Figure 6.1 above provides an overview the classic three stage architecture [1], whereby the aforementioned NLG tasks are organised into three modules: Text Planner, Micro Planner, and Linguistic Realiser. There are other types of architectures, namely the integrated, interactive, revisionist or blackboard approach [3]. Other approaches to NLG may follow an AI planning style or data-oriented approach i.e. statistical, stochastic or machine learning [3]. In some systems, NLG tasks may also split across modules and their organisation will vary. Recent data-oriented approaches, in particular deep learning ones, tend to be integrated end-to-end systems [4].

More recently, there have been major advances in the field of NLG since the introduction of Open AI's[3] GPT3 third generation Generative Pre-trained Transformer. When presented with only a few examples of text, GPT3 systems are capable of generating large volumes of relevant fluent and cohesive text. However, GPT3 models are based on deep learning and represent some of the largest neural networks ever produced. The most successful models are of course based on well-resourced languages such as English and French thanks to wide availability of training data (written digital content).

6.4 What has been done for Irish?

To date, there is no known standalone NLG system or application tool for Irish. However, use cases and potential applications (as discussed above) exist and will continue to present themselves as technology uses advance in our daily lives. This area of Language Technology can undoubtedly draw on successful work and proven techniques from other areas of the field.

Work in Machine Translation (Chapter 11) and Parsing (Chapter 5) feeds naturally into the field of NLG. When parsing textual data as into into other representations the parser is deriving 'knowledge'. Language generation is not simply the reverse of parsing, despite how counter-intuitive this may seem. Many important problems in language understanding (i.e. linguistic ambiguity) are not challenges for NLG [2]. Nevertheless linguistic grammar formalisms used for parsing natural language lend themselves naturally to generation tasks such as linguistic realisation notably. Another approach is to treat NLG as a "parsing" problem, using probabilistic context-free grammar (CFG) formalisms, considering NLG as the `inverse' of semantic parsing, whereby (e.g. inputs such as snippets of weather data) are expanded using CFG rules which use corpus derived probabilities to control "choice" [9]. More recently, treebanks have also been used in the development of NLG systems. Similarly, when translating between a source and target language the MT engine must, at some point, generate output in the target language which is correct according to the 'knowledge' it has derived from the source input. In this way, MT work on decoding (in the data-driven approach) and post-generation (in the rule-based approach) can form the basis for initial work on Irish NLG, and as R&D brings new improvements to those areas, these improvements can benefit NLG systems.

6.5 Recommendations

- One recommendation that would have critical impact in the Irish NLG space would be to adapt the widely popular rule-based surface realiser SimpleNLG [11] for Irish. The tool has become increasingly popular in industry and academia has a first line approach to generation tasks where there is a deficit of data-to-text resources for a given language with implementations now existing for German, Spanish and Italian. This is widely the case in NLG, hence the majority of newer, data-driven (deep learning) approaches only having been applied to English. Rule-based outputs can be also be used to boostrap data-driven methods.
- A SimpleNLG localised to Irish would drive basic NLG research for the implementation of automated broadcasts, reporting or public service announcements on behalf of the State, such that these services will no longer just serve the English-speaking population in Ireland.
- This is closely linked to the Information Retrieval recommendations made in Chapter 12 with respect to generation of snippets of document summaries for search results (extractive summarisation).
- NLG research is required for developing systems that can automatically summarise text in digital content (e.g. online news articles) in a form that is easier or quicker to read (abstractive summarisation).
- In order for the Irish-speaking community to

[3] https://openai.com/api/

avail of the increased use of chatbots in online services, extensive research is required not only for simple interactions but also for systems with an advanced level of emotional awareness.

- Intelligent Computer Aided Language Learning systems (Chapter 14) will rely on the availability of high quality and reliable NLG tools. Cross-discipline research is therefore required.
- Spoken dialogue systems (Chapter 10) rely on the NLG systems to produce human quality speech (in text format first).

References

[1] Reiter, E. & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

[2] Gatt, A. & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core asks, applications and evaluation. *Journal of Artificial Intelligence Research 61*, 65–170.

[3] Dale, R., Scott, D. & Di Eugenio, B. (1998). Introduction to the special issue on natural language generation. *Computational Linguistics*, *24*(3), 346-353.

[4] Novikova, J., Dušek, O. & Rieser, V. (2017). The E2E dataset: New challenges for end-to-end generation. [arXiv preprint arXiv:1706.09254].

[5] Gardent, C., Shimorina, A., Narayan, S. & Perez-Beltrachini, L. (2017). The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation* (pp. 124-133).

[6] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. & Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 178-186).

[7] Goldberg, E., Driedger, N. & Kittredge, R. (1994). Using Natural-Language Processing to Produce Weather Forecasts. *IEEE Expert 9*(2), 45-53.

[8] Reiter, E., Sripada, S., Hunter, J., Yu, J. & Davy, I. (2005). Choosing Words in Computer-Generated Weather Forecasts. *Artificial Intelligence 167*, 137-69.

[9] Belz, A. (2008). Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-Space Models. *Natural Language Engineering 14*, 431-55.

[10] Sripada, S., Burnett, N., Turner, R., Mastin, J. & Evans, D. (2014). Generating A Case Study: NLG meeting Weather Industry Demand for Quality and Quantity of Textual Weather Forecasts. In *Proceedings of INLG 2014*.

[11] Gatt, A., & Reiter, E. (2009). SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 90-93).

LCT 640 TS
lin
40 Hz
80 Hz
160 Hz
0 dB
-6 dB
-12 dB
-18 dB

Chapter 7
Speech Models for Irish

7.1 What is a speech model?

A model is an approximation of a process that captures its essential features. Of interest here are models of the speech processes to provide mathematical, concrete formulations that are the key to bringing our knowledge of Irish speech structure (Chapter 2.4) to bear on technology. These are *knowledge-based* models (linguistic-phonetic and speech science): they complement the *machine-learning* models that are widely used in technology (the two approaches are discussed in Chapter 1.7).

These models are directly based on the speech signal rather than text. Note that the resources described in Chapters 4–6 are text-based, and as such, they describe aspects of language structure that can be gleaned from text. The ways in which spoken language differs from written language are briefly discussed in Chapters 1.6 & 2.1. Speech models capture the structure of spoken language, including many kinds of information in spoken language that are not notated in writing. These are not abstract models, but capture rather the actual processes involved in human speech production, as described in Chapter 2.4.

7.2 Why important and for whom?

Knowledge-based (engineering) models combine scientific knowledge of speech processes with linguistic knowledge of the specific speech patterns of a particular language. The importance of linking these fields is discussed in general terms in §7.2.1. More specific elaboration is provided on models of voice-prosody (§7.2.2) and articulation (§7.2.3) and on their implications for future core technologies and applications for Irish.

7.2.1 Connecting linguistic and speech sciences

These models and the linkage of speech science with linguistic-phonetic knowledge allows us to:

Bridge phonetic-linguistic research ⟶ speech technology
Models provide a way to translate phonetic-linguistic knowledge into explicit technical forms that can be used in technology and presented through applications to the public. Without this bridge, much of the research on spoken Irish (§2.4) is not accessible and is likely to remain in the realms of academia.

Extend our knowledge of the spoken language
Models are powerful tools that can be used to analyse aspects of speech for which mainstream linguistic instruments are not adequate (see example in §7.2.2 below).

Link listener and speaker perspectives
Speech communication involves the speaker and listener perspectives (see Figure 15.1, Chapter 15) – an essential consideration for speech technologies that must produce and recognise speech. Models link these two. The model is initially based on the speaker's production. When the model is used to generate speech output, the model's features (i.e. parameters) can be manipulated in experiments, to establish the aspects crucial to the listener's perception – and to technology.

Enable more adequate core technology and applications
Models are particularly crucial for speech technologies and applications that need to go beyond the text-available dimensions of spoken language (see example in §7.2.2).

Facilitate knowledge transfer to language pedagogy/therapy
Pronunciation (the sounds and prosody) is often seen as the least successful aspect of Irish language teaching. Linguistic-phonetic knowledge of the structural patterns of spoken language is much less understood by the general public, teachers and learners, than grammar-related patterns that are clear in written language. A simple version of the model with visual feedback can make those important aspects of the spoken language, traditionally difficult to teach, amenable to language instruction and to speech therapy (examples below).

In the following sections we discuss models for the two basic processes of speech production, outlined in §2.4. Firstly, models of the voice source (the primary sound source), e.g. [1, 2] can be used to capture the full dimensions of voice prosody, i.e. the continuous modulation of melody, tone-of-voice, accentuation, and rhythm in Irish [3-6]. Secondly, articulatory models, such as [7-9] capture the shape and movements of the tongue and lips, which form the mouth filter to differentiate the voice source into the vowels and consonants of the language. Both kinds of models are important to future Irish speech technology and applications.

7.2.2 Modelling voice prosody: expression in the voice

Voice prosody carries many dimensions of meaning in spoken communication, in the way we say the words rather than the words themselves (§2.1). There is little notation of prosody in written text. Discussing the full role of voice prosody in carrying meaning in spoken language cannot be covered here, and we focus here on one vital aspect – how we express our feelings and our attitude to the listener by modulating our tone-of-voice.

The best current synthetic voices sound very natural and are ideal for reading (neutral) text aloud. However, they lack the expressive, affective prosody which imparts much of the meaning in our conversations. Interactive spoken dialogue systems (Chapter 10) will in the future be widely used – systems that 'listen' to us and 'answer' us, emulating human conversation. In such systems, without affective modulation of the voice, the synthetic speech can sound monotonous, unengaging and inappropriate.

Such interactive systems will increasingly feature in our everyday environment (Chapter 13) and hold particular promise for applications in Irish-language teaching, e.g. educational games, where learners engage in conversation with virtual characters that listen and talk to them (Chapter 14) and where expressive voices are needed. A lack of expressive possibilities in synthetic speech is especially problematic for those with who cannot speak, for whom the synthetic voice is their own: current technology gives them a voice – but does not allow them to modulate it to express themselves fully (see Chapter 15).

This aspect of prosody is not well understood and has been neglected by linguists because it is technically difficult to measure with standard linguistic tools. Modelling offers a way of capturing how varying the voice quality signals the speaker's emotions and attitudes [10, 11]. The models can be deployed in technology, e.g. in synthesis [12, 13] to allow control of the quality and emotional colouring of synthetic speech.

There are implications for all speech technologies, not only speech synthesis. Recognition systems (Chapter 9) currently recognise words (which are written) but do not recognise the affective interpersonal dimensions of the spoken language. There is a growing realisation of this important gap, and increasingly, research towards applications that detect the affective prosody – e.g. in commercial applications that deploy recognition systems, or in health applications. Voice prosody is a key indicator of a person's health and particularly mental health, and systems to detect it are being developed to monitor patients and ensure timely therapeutic intervention. Dialogue systems (Chapter 10) need not only to speak with expressive voices, but also to pick up on the affect in the user's speech.

For language learners, acquiring native-like prosody is challenging, and this aspect is difficult to teach, especially as it is not visible in the written form. However, a simple model of prosody with visual feedback can make this important aspect of the spoken language more amenable to explicit instruction. Note that these models can be adapted to allow for dialect-specific prosodic patterns (discussed in §2.4). Such models are also valuable aids in speech therapy.

7.2.3 Models of articulation

Articulatory models capture the movement of the speech organs, mapping it to the acoustic features of the sounds of the language. This is of particular interest for Irish, given the complex articulation of contrasting slender and broad consonants, very different to those of English. For example, the pronunciation of the first consonant in the Irish words /lˠo:nˠ/ (*lón* 'lunch') and /lʲo:nˠ/ (*leon* 'lion') is very different, and both differ from that of the English /lo:n/ 'lone' (see Chapter 2.4.2).

The articulatory modelling needed for technology was limited in the past due to the technical difficulties of obtaining high quality, high speed data, but advances in fMRI technology are leading to rapid development in this field [14, 15]. Articulatory models are likely to feature in future technologies, such as articulatory synthesis [9] or speech recognition where articulatory features assist the decoding.

Articulatory modelling has useful applications in education and speech therapy. Many learners of Irish struggle with the articulation of the consonants. Failure to master the broad/slender contrast can affect not only the learner's pronunciation but also the acquisition of literacy and certain aspects of grammar (Chapter 2.4.2 and Chapter 2.7.3). Simplified articulatory models with visual feedback can be used

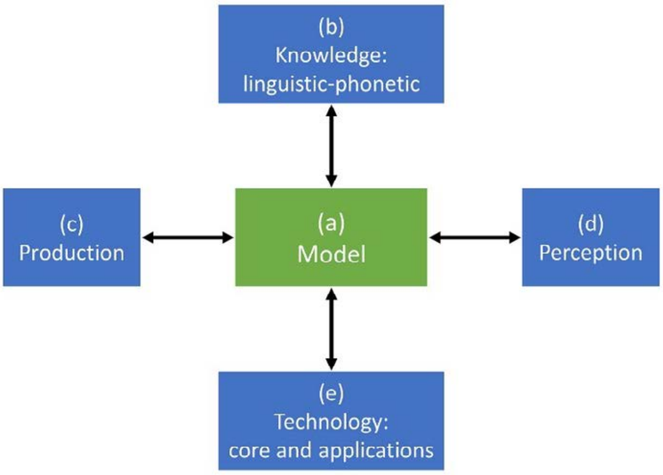


Figure 7.1: The model as the hub that brings linguistic-phonetic knowledge into speech technology, allowing a continuous cycle of development

for pronunciation training [16] and could be of great value in teaching Irish pronunciation. These would likewise be helpful in speech therapy for those with articulation difficulties. Visual feedback makes aspects of pronunciation intuitively obvious, that are otherwise difficult to put across and not obvious in writing.

7.3 How do models work?

There are many kinds of models that can be used, and given their highly technical nature, it is beyond the scope of this document to explain the working of a specific model. However, Figure 7.1 illustrates how the process of modelling works as the essential hub in a cyclical process of development that spans the divide between linguistic-phonetic research and technology:

The data for ***the model (a)*** comes initially from ***linguistic-phonetic knowledge (b)*** (§2.4). The model's output is tested against ***speaker's production (c)*** and ***listeners' perception (d)***. This allows improvements to the model and extends our linguistic understanding. ***Implementation in technology applications (e)*** yields test platforms, where users' feedback identifies the model's strengths and limitations. This leads to further improvements to the model, extends our knowledge and enables further enhancement of technology.

7.4 What has been done to date?

Voice Prosody Modelling: A parametric model of the voice [1, 2] is being used to capture the tone-of-voice dimension of Irish prosody [5, 13, 17] within the Róbóglór[1] project at Trinity College, Dublin. This work builds on earlier analyses of speakers' production and listeners' perception [6, 10, 11] and adds to the linguistic descriptions of the melodic patterns in Irish dialects (see Chapter 2.4.3). Ultimately, this will provide a more holistic picture of Irish dialect prosody that will include the affective, expressive dimension. It is currently guiding exploration of voice parameters that can be controlled in synthetic speech and used in experiments with the ABAIR Irish synthetic voices[2] [13].

Articulation Modelling: to date, although a (limited) number of articulatory studies have been carried out on the Irish consonants (see §2.4.2), until recently it was not possible to capture the kind of data needed for articulatory models that would be optimal for technology.

[1] The Róbóglór project is supported by the Department of Tourism, Culture, Art, Gaeltacht, Sport & Media.

[2] Synthetic voices for the Irish dialects are available at www.abair.ie. The ABAIR project is supported by the Department of Tourism, Culture, Art, Gaeltacht, Sport & Media.

7.5 Future work

As the kind of knowledge-based modelling discussed here is at the intersection of linguistic-phonetic research and technology development, it calls for an interdisciplinary approach. On the one hand, close collaboration with phonetic research will ensure that modelling supports acquisition of linguistic data, which in turn feeds the testing and refining of the models. On the other hand, modelling research needs to progress alongside speech technology building as referred to in Parts II and III. The methodology used to build a specific technology may determine how the models are formulated and how they can be incorporated. For example, in speech synthesis (Chapter 8), a variety of speech engines can be used: some would allow direct implementation of the kinds of knowledge-based models described here, while others require more indirect methods for incorporation.

One of the targets for voice prosody modelling will be to extend the prosody in the currently available Irish synthetic voices to incorporate an affective, expressive dimension (see Chapter 8). Although modelling the full nuancing of human prosody remains an elusive goal, incremental approximations towards this goal can be achieved in limited domain systems for specific applications (e.g. interactive educational games, communication aids) to enhance their appeal and acceptability for users. Over time, one would aspire to extending the technologies and range of applications where speech technology can be used effectively for Irish.

With the advances in using fMRI imaging for articulatory modelling, this area is opening up for speech technology [14, 15], and it is now possible to obtain clear, usable data on articulation and on how it maps to the acoustic signal. To advance this goal, a collaboration of Irish language specialists with international laboratories leading in this field is recommended in the first instance, to jointly develop the extensive datasets needed to model the articulation of Irish consonants and vowels. Further development of research expertise here would be needed in the longer term, to permit more extensive coverage of dialects and extend the range of applications for the public.

Sophisticated models will greatly deepen our understanding of the language and enable us to harness new paradigms that emerge in speech technology. Exploiting such models in teaching and therapy should enable intervention in areas which are important but have been inaccessible to date.

Although modelling is exploratory fundamental research which draws on the state-of-the-art in the field, it is important nonetheless to target specific applications from the outset where such models are likely to have most impact. Early, proof-of-concept systems will provide evaluation testbeds. This helps ensure the research outputs are used to enrich user-applications, maximising their potential benefits to the Irish-language community, even if the full impact emerges incrementally over some years.

7.6 Recommendations

As discussed above, modelling research needs to proceed alongside phonetic-linguistic work, providing the data and continuously testing the models. The close collaboration with core technology and application building is also crucial, to ensure that the research outputs are reaching the public, and that user's feedback leads to continuous improvement to models. Therefore, recommendations for this area of research are included in the recommendations for basic research on spoken Irish (Chapter 2.4), for core speech technologies (Chapters 8, 9 & 10) and for applications that are based on them.

A priority will be to foster interdisciplinary teams with a focus on Irish, encompassing phoneticians, speech scientists (engineers) skilled in modelling as well as experts in building the latest speech technologies. As explained in Chapters 1 and 16, establishing this type of human infrastructure will be an essential foundation for this sector into the future. Core recommendations are therefore:

- **Develop models of voice prosody** in Irish dialects, to capture the main features that impart the linguistic and expressive meaning of the spoken language. This requires collaboration of phonetic and speech engineering science.

- **Develop models of Irish articulation**, focusing on the consonantal contrasts and their interaction with vowel articulation. This work will crucially require collaboration with foreign experts in the technologies, as well as a local collaboration of phonetics and speech engineering science.

- **Explore how these models can be incorporated** into core Irish speech synthesis and recognition technologies. As the means of implementation depend on the synthesis and recognition engines used, researchers from the different groups need to work together.

- **Build proof-of-concept applications** that incorporate the models developed in the first two recommendations above. It is recommended that the primary focus for such applications would initially be on applications for Irish-language teaching, speech therapy and applications for people with disabilities. For both prosodic and articulatory modelling, applications with simplified visual feedback should be explored. This work would require close collaboration with the targeted users, their teachers, therapists etc.

References

[1] Fant, G., Liljencrants, J. and Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR (Speech, Music and Hearing, Royal Institute of Technology, Stockholm) 26*(4), 1-13.

[2] Fant, G. (1997). The voice source in connected speech. *Speech Communication 22*, 125-139.

[3] Gobl, C., & Ní Chasaide, A. (2010). Voice source variation and its communicative functions. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2 ed.). Oxford: Blackwell Publishing Ltd., pp. 378-423.

[4] Iseli, M., Shue, Y.-L., Epstein, M. A., Keating, P., Kreiman, J., & Alwan, A. (2006). Voice source correlates of prosodic features in American English: A Pilot Study. In *Proceedings of Interspeech 2006* (pp. 2226-2229).

[5] Ní Chasaide, A., Yanushevskaya, I. & Gobl, C. (2015). Prosody of voice: declination, sentence mode and interaction with prominence. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

[6] Yanushevskaya, I., Gobl, C. & Ní Chasaide, A. (2017). Cross-speaker variation in voice source correlates of focus and deaccentuation. In *Proceedings of Interspeech 2017* (pp. 1034-1038).

[7] Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America* 53, 1070–1082.

[8] Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W.J. Hardcastle & A. Marchal (Eds), *Speech production and speech modelling*. Boston: Kluwer Academic Publishers, pp. 131–149.

[9] Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS One 8*(4).

[10] Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2), 189-212.

[11] Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2018). Cross-language differences in how voice quality and f0 contours map to affect. *Journal of the Acoustical Society of America 144*(5), 2730-2750.

[12] Sorin, A., Shechtman, S. and Rendel, A. (2017). Semi Parametric Concatenative TTS with Instant Voice Modification Capabilities. In *Proceedings of Interspeech 2017* (pp. 1373-1377).

[13] Murphy, A., Yanushevskaya, I., Ní Chasaide, A. & Gobl, C. (2021). Integrating a Voice Analysis-Synthesis System with a TTS Framework for Controlling Affect and Speaker Identity. In *Proceedings of the 32nd Irish Signals and Systems Conference.*

[14] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S. & Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *Journal of the Acoustical Society of America 136*(3), 1307-1311.

[15] Ramanarayanan, V., Tilsen, S., Proctor, M., Töger, J., Goldstein, L., Nayak, K. S., & Narayanan, S. (2018). Analysis of speech production real-time MRI. *Computer Speech & Language 52*, 1-22.

[16] Suemitsu, A., Dang, J., Ito, T., & Tiede, M. (2015). A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *Journal of the Acoustical Society of America 138*(4).

[17] Murphy, A., Yanushevskaya, I., Ní Chasaide, A. & Gobl, C. (2018). Voice Source Contribution to Prominence Perception: Rd Implementation. In *Proceedings of Interspeech 2018* (pp. 217-221).

[18] Ní Chasaide, A., Yanushevskaya, I., Kane, J. & Gobl, C. (2013). The voice prominence hypothesis: the interplay of F0 and voice source features in accentuation. In *Proceedings of Interspeech 2013* (pp. 3527-3531).

Part II

Core technologies

k
i
h
g
Plato
g
f
e
d
c

Chapter 8
Speech Synthesis

8.1 What is speech synthesis?

Speech synthesis is the generation of speech using a machine. It typically involves a system that converts text automatically into synthetic speech, called a text-to-speech (TTS) system. Nowadays, synthetic speech is widely heard, e.g. on your computer, from the virtual assistant on your phone, in the lift, in computer games etc. For many with disabilities it may be the voice with which they communicate. It can sound almost indistinguishable from the natural voice, or it can sound fairly unnatural and robotic, like the synthesiser which was used by Stephen Hawking.

As with all speech technologies for Irish, as there is no single spoken standard, it is necessary to provide for the different dialects.

8.2 Why and for whom is it important?

For any endangered language, speech synthesis is an essential technology as it places the living, spoken language centre stage in the digital world: by embedding speech synthesis into applications, we extend the everyday domains where we use and are exposed to spoken Irish.

Why and for whom speech synthesis is important can therefore be quite different for the minority languages compared to a widely spoken language like English. The implications for Irish can be illustrated through the uses of TTS voices already developed for the Irish dialects by the ABAIR initiative[1] in Trinity College Dublin (specific applications are discussed in §8.4).

A public utility for the wider Irish-language community, including the diaspora: considerable traffic on the ABAIR website[2], which provides TTS voices for Irish dialects, demonstrates widespread public demand. The large numbers from outside Ireland highlight how dispersed the Irish language community is. User feedback emphasises the importance of hearing the spoken, native-speaker version of written language (more below).

Disability/access: speech synthesis is a basic need for many with disabilities. A screen-reader with synthetic speech output is essential to allow participation of the visually impaired in Irish-medium education and in the online social and cultural world of Irish. For those who cannot speak (e.g. many autistic people, people with cerebral palsy, Parkinson's etc.) speech synthesis is central to communication and inclusion. It is also an essential aid for many with dyslexia. These aspects are discussed in Chapter 15.

Education: for a minority language, the potential impact for language learners is enormous. Most learners of Irish have limited access to native speaker models of the language: speech synthesis brings a 'virtual native speaker' into the classroom and the home – a powerful learning aid to get the sounds of the language on your ear and to develop speaking and aural skills. Linking speech to text is particularly helpful for literacy acquisition, given the opaque sound-to-letter correspondences of Irish (see Chapter 2.4.2 & Chapter 2.7.3). Speech synthesis opens the door to powerful interactive educational platforms and games, which could greatly impact language teaching and motivate learners (see Chapter 14).

Language maintenance and documentation: a synthetic voice preserves a virtual speaker for posterity – of itself a powerful form of documentation and preservation [1, 2]. This offers the local community of an endangered dialect very tangible support in its efforts to revive and maintain it – especially when coupled with the language resources developed as part of the building process (§8.3.1) and the types of applications it allows (§8.4.2). Even for a dialect which has been lost, with suitable recordings, it may be feasible to resurrect a virtual speaker.

Research on the Irish dialects: as building TTS systems involves developing linguistic resources (Chapter 2), corpora (Chapter 3) and speech models (Chapter 7), it helps extend our knowledge of the Irish dialects and provide valuable materials and new methodologies for future researchers.

[1] ABAIR is supported by An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán
[2] www.abair.ie

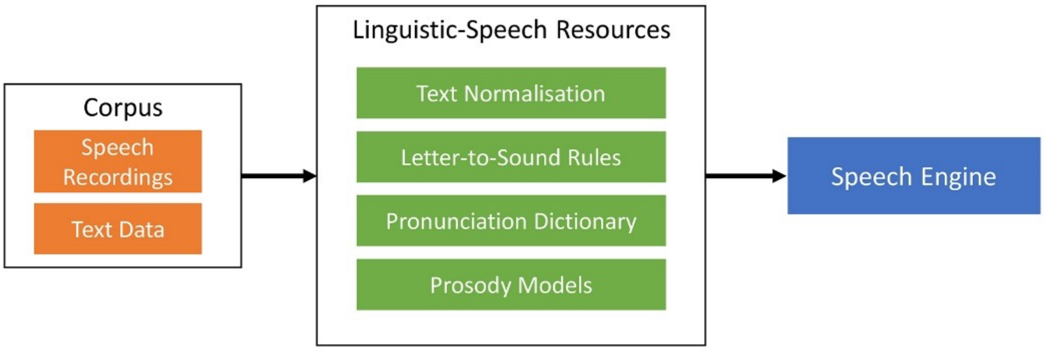


Figure 8.1: The modular TTS system involves the development of speech corpora, linguistic-speech resources and a speech engine

Combining synthesis with other technologies, such as speech recognition (Chapter 9) to enable interactive dialogue systems (Chapter 10), greatly extends the scope and impact of applications. In the not-too-distant future, speech-based interaction with machines and robots will be commonplace as well as applications such as speech-to-speech translation, where speech synthesis and recognition are combined with machine translation.

8.3 How does it work?

There are different ways in which speech synthesis systems can be developed [3]. They can be grouped into two very different types, (a) modular systems involving a speech corpus, linguistic components and a speech engine, and (b) End-to-End (E2E) systems which dispense with many explicit linguistic components, and where deep learning is used to establish the correspondences between recorded speech data and matching text. The latter has been shown to produce high-quality speech output in languages such as English. Note that these E2E systems require vast amounts of training data and computing resources.

8.3.1 Modular systems

Figure 8.1 illustrates the three modules of such TTS systems, where the linguistic-speech components mediate between the speech corpora and the speech engine.

Corpora

In these systems, the synthetic voices are based on the voice of a single representative speaker of a language or a particular dialect, for whom a corpus of high-quality recordings is made. In this corpus, the sounds are identified, phonetically transcribed and aligned to the acoustic signal. Corpus design is important: as sounds are heavily influenced by the preceding/following sounds and by the prosodic context (§2.4), the corpus should contain instances of each sound in every context. The *ARCTIC* corpus [4] is designed in this way for English: given the very large consonantal inventory of Irish, a correspondingly very large corpus is required. Corpora need to be developed for each dialect, with full coverage of its sounds and ideally recorded with dialect-appropriate materials.

Linguistic & Speech Resources

The TTS system uses linguistic resources and speech models (Chapter 2.4 & 7), for a given dialect, that enable conversion of written text and symbols into sequences of sounds, specify prosodic features and prepare the input so that it is adapted to the speech engine. The **text normalisation** component rewrites numerals, acronyms, symbols, etc. as text. The **letter-to-sound** component (§ 2.4.4) converts written forms to the sequence of sounds they represent, e.g. *buíon* ⟶ /bˠiːnˠ/ 'band'. Irregular forms not conforming to the letter-to-sound rules are stored in a **pronunciation dictionary** for the dialect e.g. *(ní) bhfaighfidh* /wˠːi/ or /vˠaɪgʲ/ *(not) get*, future tense, in Donegal or Kerry dialects respectively (see Chapter 2.7.3). **Acoustic models** (capturing the acoustic features for each sound) are constructed for every sound in every context. **Prosody models** capture the patterns of melody, accentuation, voice quality etc. (Chapter 2.4.3 and 7).

The choice and organisation of components depend on the speech engine used (see below). *All* these components must be appropriate to the individual dialect.

[3] These are knowledge based models and they are explained in Chapter 7.

Speech Engine

The speech engine takes the string of sounds (typically, acoustic models with prosodic specifications) and generates the speech output. There are different kinds:

Parametric, formant (source-filter) synthesis generates a speech signal based on acoustic models of the *voice (source)* and *vocal tract (filter)* characteristics of each sound (Chapter 2.4 and 7)[3]. Early synthesisers were of this type, and they are still widely used in assistive technologies for people with disabilities (most famously by Steven Hawking). Depending on the sophistication of the modelling, the speech can sound robotic or very natural.

Unit selection synthesis doesn't generate speech from acoustic specifications but rather searches for and concatenates sounds (or sound fragments) taken from an appropriate context in a large spoken corpus [5]. It can sound very natural, but it can also 'fail' spectacularly – for a variety of reasons, such as inadequate coverage of sounds/contexts in the corpus, errors in the transcription, etc. As a large corpus is stored and searched to generate speech output, it is too slow for applications needing high speed output (e.g. screenreaders for the blind), and too 'bulky' to incorporate into small devices, such as phones.

Statistical Parametric Speech Synthesis (SPSS) *generates* speech from the stored acoustic models for every sound in every possible context, obtained by machine-learning techniques (see Chapter 1.7), using **HMMs** [6, 7] and, increasingly employing deep neural nets, **DNNs** [8, 9]. They provide rapid, stable and increasingly high-quality speech output, and are widely used.

Articulatory speech synthesis emulates human speech production, generating speech directly from mechanical models of the source and filter (Chapters 7 & §2.4). Though not (yet) commercially viable, this could be the technology of the future, as advances in other technologies now permit capture of the rich production data (voice source and filter) needed for articulatory synthesis. Being based directly on models of human production (voice and filter) this approach could have many advantages in quality and flexibility. It would permit a single synthetic 'speaker' to be transformed into multiple 'speakers' (men, women, children) by rescaling the dimensions of the articulatory model. It could also enable control of voice prosody to resemble the expressive prosody of human speech (see below and Chapter 7).

8.3.2 End-to-End systems

Speech synthesis technology is developing rapidly. Deep neural networks are increasingly being used in End-to-End systems, and feature very large scale spoken corpora with numerous speakers, and no (or less explicit) use of linguistic resources [10 - 13]. While this approach works well for a language such as English where vast corpora with large numbers of speakers are available, this is not the case for lesser-resourced languages. However, these systems will become increasingly feasible for Irish, when very large multi-speaker corpora for the various dialects become available.

Although offering great promise, there are currently a number of issues with E2E systems for Irish and reasons why an approach which retains linguistic information is still needed. One issue concerns errors in the output of the E2E system; these are very difficult to diagnose and remediate without explicit access to the linguistic components. A further issue concerns the footprint and the computational power that E2E systems require. At the moment, this presents an obstacle for many applications, which need to run on small devices with limited memory and must be packaged to run offline. But perhaps most importantly of all, the embedded linguistic knowledge is critical to many of the applications from which the impact of Irish TTS derives for the language community (see §8.4 below).

In the case of Irish, both modular and end-to-end systems are important. Hybrid systems that can optimise the advantages of both will undoubtedly also be important.

8.4 What has been done to date?

This is an area of considerable research activity for Irish. TTS systems for the Irish dialects have been developed in tandem with applications needed by the language community. These two aspects are discussed in the following sections.

8.4.1 Irish text-to-speech systems

TTS systems for the three main dialects have been developed as part of the ABAIR project, based on native speakers' speech, and these are freely available on the project's website[4]. These include Gaeltacht voices for Ulster (Gaoth Dobhair, female), Connaught (Conamara, male), Munster (Corca Dhuibhne, female and male) and further voices are under development. Various speech synthesis methods (and speech engines) are deployed. Users can choose the dialect/speaker of preference and can select the speech synthesis modality/engine that

[4] www.abair.ie.

best suits the requirements of a particular application. Recently, Microsoft made a male and a female Irish synthetic voice available (based on non-Gaeltacht speakers). A Scottish company, Cereproc, also offers an Irish Ulster (female, non-Gaeltacht) voice.

Current ABAIR research is focusing on extending and enriching the diversity of speech corpora and synthetic voices. Research on E2E systems for the Irish dialects is ongoing, with an emphasis on hybrid systems which retain critical linguistic components.

8.4.2 Speech synthesis-based applications

A number of applications incorporating the ABAIR voices have been developed in response to demand from (and in collaboration with) individuals and community groups. Many of the following are hosted on the ABAIR website, while some are at prototype stage.

A Webreader[5]: this allows the content of webpages to be read aloud. The user chooses the dialect, the speech engine and the speed of the speech output. The text being spoken may be magnified and highlighted as it is being read out.

An Android phone app[6] is available, featuring the Conamara (HMM) voice. Further dialects are currently being added.

A Screenreader (NVDA plugin) for the visually impaired[7]: a plugin for the free, open source NVDA screen-reader provides a choice of ABAIR voices. It reads out text and all information needed to navigate computer functions, programmes and applications. The user selects the dialect and speech engine and controls the speed of speech output. This system is also compatible with the Liblouis Braille system, thus allowing simultaneous Braille and speech output [14] (more in Chapter 15).

DAISYbooks[8]: the ABAIR voices are used in voice-enabled multimodal textbooks for visually impaired schoolchildren.

Speech generating AAC system: this is a system that allows a non-verbal person to communicate by selecting a series of words/images which are then spoken out by the synthetic voice. An initial prototype system for Irish has been configured [15, 16] and will need considerable further development (see Chapter 15).

Educational applications: the ABAIR voices are being used in Irish language-learning platforms and interactive games currently under development. These include an *iCALL* (intelligent computer-assisted language learning) platform *An Scéalaí*[9] [17], which aims to develop speaking and listening skills in parallel with writing and reading skills. For early learners, an interactive multimodal game *Lón don Leon* [1, 15] targets the development of phonological awareness and early literacy. The interactive educational game *Digichaint*, [18], and the chatbot *Taighdhín* [19] have been developed to proof-of-concept level. The ABAIR voices are also integrated into educational platforms such as *Edcite*[10]. (See discussion of educational applications in Chapter 14).

These applications and the interactive games highlight the need for multiple characters and children's voices (in the individual dialects) - for example, to populate the casts in games and for adequate access/disability/ educational applications. They also highlight how future synthetic voices will need the capacity to generate speech with the kind of expressive voice prosody of human spoken interactions (see Chapters 2.4.3 and 7). Early-stage voice-prosody modelling towards these goals is being carried out as part of the *Róbóglór*[11] project at Trinity College Dublin.

8.5 Future work

Building on the current provision of synthetic speech, a priority for future development is to ensure that systems are suited to the wide variety of envisaged future users and applications that will maximise its positive impact in the community. Extension of synthesis to further dialects, including the most endangered is important. Speech technology is rapidly changing, and the continuous extension, evaluation and upgrading of all aspects of TTS will be needed, keeping abreast of novel techniques as the field advances.

[5] The ABAIR Webreader can be downloaded at https://www.abair./webreader

[6] The ABAIR Android App: https://abair.ie/android_app. There is also information about the app here: https://abair.ie/products/abairapp/index.html

[7] Plugin for Irish speech output in the NVDA Screenreader can be downloaded from https://www.abair.ie/nvda. This work was also supported by the National Council for the Blind of Ireland (NCBI) and An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG).

[8] DaisyBooks were prepared in collaboration with *Childvision in Ireland*. Further information on DaisyBooks is available at: http://www.daisy.org/daisypedia/daisy-digital-talking-book.

[9] Beta version of An Scéalaí can be found at https://www.abair.ie/scealai. This work is also supported by An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG).

[10] Edcite web page at: https://www.edcite.com/

[11] The Róbóglór project is funded by the Department of Tourism, Culture, Art, Gaeltacht, Sport & Media

Continued ***interdisciplinary*** research is recommended. (as an Irish version) A close linkage of synthetic speech development with provision of *linguistic-phonetic and speech resources* for the dialects (Chapters 7 & §2.4) will be important. Such resources, developed for synthesis, are reusable in further technologies (e.g. speech recognition) and key to properly focused and dialect-appropriate applications, particularly for education, disability and access.

The collaborative development of speech-based *applications* alongside core TTS systems is also important. It is primarily through the applications that the core technology reaches the language community. This approach also supports the core technology, as the applications provide the ideal platform for evaluating the synthetic voices (see for example [20]). Urgently needed applications (such as the screenreader and the AAC system) further guide the priorities in core technology development. (For example, the high-speed speech output needed in the screenreader prompted ABAIR's initial development of SPSS speech engines). In the future, as speech synthesis and recognition (Chapter 9) are coupled, speech synthesis priorities will focus more and more on 'intelligent' interactive systems (Chapter 10) for specific user groups. These kinds of considerations need to be borne in mind in setting the research agendas.

It is increasingly clear that investment in software support is essential to ensure that Irish speech synthesis is embedded in modern speech-based devices and applications, that they continuously upgraded as technology evolves and as operating systems change.

8.6 Recommendations

Recommendations for future development include:

- **Extending synthesis provision:** dialects, children's voices, bilingual synthesis to accommodate a wide range of users and applications. The following is necessary:
  - ***Further dialects*** need to be provided for, including those that are the most endangered.
  - ***Linguistic resources*** and corpora are needed for each dialect, based on appropriate materials for the dialect and speaker, with full coverage of the dialect's sound combinations etc. Very large multi-speaker recordings of high quality will be needed to yield a Voicebank for each dialect. As targeted applications evolve, so too will the types of spoken corpora needed.
    - ***Children's voices of differing ages, and a wide choice of voices:*** children's voices are a key requirement for future educational platforms, and for disability/access applications. Between young and old, a choice of voices with different intrinsic voice qualities will be desirable.
  - ***Bilingual synthesis*** will be needed that can deal with the frequent code switching (English words/phrases in Irish text) and with bilingual websites.
- **Continuous development of speech engines and synthesis methods:** Irish TTS will need to evolve in line with the rapid developments in speech synthesis technology. It will also be important to ensure that the resources recommended above and the innovations recommended below are harnessed effectively as the technology changes.
- **Expressive voices and voice transformation**: controlling the voice and prosody in synthetic speech (Chapter 7) is important to future applications that will increasingly feature interactive dialogue systems where we converse with virtual agents, chatbots and characters in games (Chapters 10, 13, 14, 15). This requires:
- ***Expressive voice:*** while current systems are intelligible and sound natural, they do not allow control of the voice to express the feelings and interpersonal information (Chapters 2.4.3 and 7.3) conveyed in human interactions. Putting expression into the voice is critically important for those users with disabilities, whose synthetic voice is their own voice.
  - ***Transforming the voice***: controlling voice parameters will also allow the generation of multiple 'speakers' from a single synthetic voice, as needed to populate the host of characters in games. It would also permit fine tuning of the synthetic voice for individuals with disabilities, in keeping with their sense of identity. When recordings of a great many speakers for each dialect becomes available ('Voicebanks'), it would also be possible for disabled users to select a voice to which their synthetic voice could be tuned.
  - **Exploiting synergy with speech recognition:** there is an increasing overlap in some of the methodologies, resources and expertise needed to develop speech synthesis and recognition (see Chapter 9). Although requirements can be very different, a joint approach e.g. in designing and developing corpora and resources strengthens both areas.
- **Developing prototype applications into full systems**: some synthesis-based applications discussed in §8.4.2 above, such as interactive educational games, have been developed to prototype stage. These and other applications need to be developed into full systems, tested and disseminated to schools, to specific user groups and to the public. This entails deep

interaction with the language community and professionals in the fields of education, disability etc. Dissemination will require close cooperation with State institutions to ensure applications are widely available, used and supported.

- **Testing:** continuous testing of synthesis and of synthesis-based applications is required.
- **Packaging, interfacing and maintenance:** to be ubiquitous, the synthetic voices need to be packaged for use in different devices with differing operating environments. Interfaces are needed to embed the voices in specific hardware devices and peripherals. All interfaces and applications need continuous updating in the rapidly evolving technology world and maintained to keep up with changes in operating systems. As systems become widely used, user support will be essential.

References

[1] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy, A., Barnes, E. & Gobl, C. (2019). Leveraging phonetic and speech research for Irish language revitalisation and maintenance. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 994-998).

[2] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., & Murphy, A. (2015). Speech technology as documentation for endangered language preservation: The case of Irish. In *Proceedings of the 18th International Congress of Phonetic Sciences*.

[3] Taylor, P. (2009). Text to Speech Synthesis. Cambridge University Press.

[4] Kominek, J., & Black, A. W. (2004). The CMU Arctic Speech Databases. [Paper presentation]. *SSW5-2004*.

[5] Hunt, A. & Black, A.W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP, Volume 1* (pp. 373-376).

[6] Zen, H., Tokuda, K. & Black, A. (2009). Statistical Parametric Speech Synthesis. *Speech Communication 51*(1), 1-23.

[7] Tokuda, K., Tomokie, T. & Yamagashi, J. (2013). Speech Synthesis Based on Hidden Markov Models. In *Proceedings of the IEEE* (pp. 1234-1252).

[8] Watts, O., Henter, G. E., Merritt, T., Wu, Z. & King, S. (2016). From HMMs to DNNs: Where do the improvements come from?. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

[9] Zen, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

[10] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing* (pp. 4779-4783).

[11] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, Robust and Controllable Text to Speech. *Advances in Neural Information Processing Systems 32*.

[12] Łańcucki, A. (2020). FastPitch: Parallel Text-to-speech with Pitch Prediction. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing* (pp. 6588-6592).

[13] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. http://arxiv.org/abs/2006.04558

[14] McGuirk, R. (2015). *Exploration of the Use of Irish Language Synthesis with a Screen-Reader in the Teaching of Irish to Pupils with Vision Impairment*. [M.Phil. thesis]. Trinity College, Dublin.

[15] Ní Chasaide, A., Barnes, E., Ní Chiaráin, N., McGuirk, R., Morrin, O., Nic Corcráin, M., & Cummins, J. (2022). Challenges in assistive technology development for an endangered language: an Irish (Gaelic) perspective. In *Proceedings of the 9th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 80-87).

[16] Barnes, E., Morrin, O., Ní Chasaide, A., Cummins, J., Berthelsen, H., Murphy, A., Nic Corcráin, M., O' Neill, C., Gobl, C. & Ní Chiaráin, N. (2022). AAC don Ghaeilge: Prototype Development of Speech-Generating Assistive Technology for Irish. In *Proceedings of the CLTW 4 @ LREC2022* (pp. 127-132).

[17] Ní Chiaráin, N. & Ní Chasaide, A., (2019). An Scéalaí: autonomous learners harnessing speech and

language technologies. [Paper presentation]. *SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education.*

[18] Ní Chiaráin, N., & Ní Chasaide, A. (2016a). The Digichaint interactive game as a virtual learning environment for Irish. In *Proceedings of EUROCALL 2016.*

[19] Ní Chiaráin, N., & Ní Chasaide, A. (2016b). Chatbot technology with synthetic voices in the acquisition of an endangered language: Motivation, development and evaluation of a platform for Irish. In *Proceedings of the 10th International Conference on Language Resources and Evaluation.*

[20] Ní Chiaráin, N., & Ní Chasaide, A. (2015). Evaluating synthetic speech in an Irish CALL application: influences of predisposition and of the holistic environment. In *Proceedings of SLaTE 2015: 6th Workshop on Speech and Language Technologies for Education.*

SONY
MIC STEREO HEADPHON
MDR-7506
Professional

Chapter 9
Speech Recognition

9.1 What is it?

Automatic speech recognition (ASR) involves the conversion of spoken utterances into text, i.e. speech-to-text, and is the companion technology to text-to-speech synthesis. Speech recognition allows the user to 'write' text by speaking to the computer. It is used for note-taking, record keeping and to speed up the process of writing. It is an essential tool for many with dyslexia, and it can assist communication for many with disabilities and facilitate their inclusion in social, educational and professional spheres. Speech recognition is also increasingly used to control electronic devices. In call centres, it is often a computer that 'recognises' our query.

Speech recognition and speech synthesis are used not only in stand-alone applications, but increasingly in complex systems like electronic 'virtual assistants' where users talk to and get information from machines (see Chapter 10). Alone or integrated with other technologies, speech recognition is becoming a feature of our everyday environment.

9.2 Why important and for whom?

Speech recognition is a particularly important technology for Irish because, like synthetic speech (and especially when linked to it) it enables applications that promote the use of the spoken, living language. It helps change how people learn Irish and how they use it in their everyday working and leisure activities. It is used in wide ranging applications such as:

Dictation systems allowing text to be composed by voice, dispensing with the need for a keyboard. This would have wide public appeal and should prove useful to professionals who work with Irish, whether as authors, administrators or public servants and all who write in Irish. Learners at all levels can benefit for writing assignments/homework – further facilitated by the addition of grammar and spell-checkers. It is an essential tool for those who can speak but may have not (yet) have mastered written Irish – invaluable for those with dyslexia, early learners or those with motor or visual impairment.

Note keeping in hands-busy, eye-busy situations: similar devices allow written record keeping when the eyes/hands are being used with other matters, such as when operating machinery, when driving, when doctors examine patients, when scientists/students conduct experiments.

Automatic subtitling of television programmes, radio etc. enables the spoken word to appear as text at the bottom of the screen in time with the programme or film. Automatic captioning of any audio signal (e.g. a public lecture) can also be very helpful to many members of the public, in education and is a critical facility for those with hearing difficulties.

Voice control of electronic devices: simple spoken commands can, directly or remotely, control devices such as the television, radio, computer, phone, sound system. One can search for programmes or apps, regulate home heating and other systems etc., overcoming the need for complicated interfaces, remote controls and keyboards.

Computer-Assisted Pronunciation Training (CAPT) is a technology derived from speech recognition, which is increasingly used in language-learning applications for the major languages. It compares learners' productions with a target pronunciation and provides corrective feedback.

Health Monitoring: although not part of mainstream recognition systems, there is a growing focus on applications which extract features such as voice prosody information relating to emotional state and attitude (see Chapters 2.4 & 7). These can provide a measure of the speaker's state of health, and particularly mental health, to enable earlier therapeutic intervention for those at risk.

Integrating ASR with speech synthesis and other language technologies, enables applications, such as:

Dialogue systems: Increasingly, we will interact by voice with machines, robots and virtual agents. Dialogue systems can offer truly interactive immersive environments, optimal for language learning (Chapters 10, 14). They will enable those with disabilities to communicate with others and interact with their environment (Chapter 15). Current developments include interactive agents to assist the elderly or disabled maintain their independence.

Speech-to-speech machine translation: while current machine translation systems entail text-to-text translation, future systems where technologies are integrated will enable a speaker of Language A to speak a message which is translated and spoken with the synthetic voice in Language B.

9.3 How does speech recognition work?

There are two main types of speech recognition system: (a) modular systems, which include an acoustic model, a language model and a pronunciation dictionary as illustrated in Figure 9.1, and explained in the following section, and (b) End-to-End (E2E) systems, which use a single component to convert speech to text. Although E2E systems have achieved state-of-the-art performance in many languages, it requires a huge amount of training data and computing resources. For low-resource languages, the modular system is more feasible, is still widely used in products and tends to yield a better performance.

A more technical explanation of speech recognition can be found in [1-3].

9.3.1 Modular recognition systems

Figure 9.1 illustrates the process involved in developing a modular speech recognition system

1 ***Preparation of corpora & pronunciation dictionaries***

Speech Corpora: very extensive speech corpora with matching text are a prerequisite. The corpora should contain numerous examples of all the sounds of the language, in all possible contexts [4]. Recordings of large numbers of speakers are needed: the more speakers included, the better the recognition outcomes.
Pronunciation Dictionary: this provides the sounds sequence corresponding to the written version of each word, with multiple entries if differing pronunciations are encountered. Note that in languages like English, pronunciation dictionaries are already available. For Irish, no suitable pronunciation dictionary is available, and they need to be constructed for this purpose. Furthermore, as there is no single spoken standard, pronunciation dictionaries are needed that reflect the pronunciation of the individual dialects.
Text Corpora: Extensive text corpora are also used.

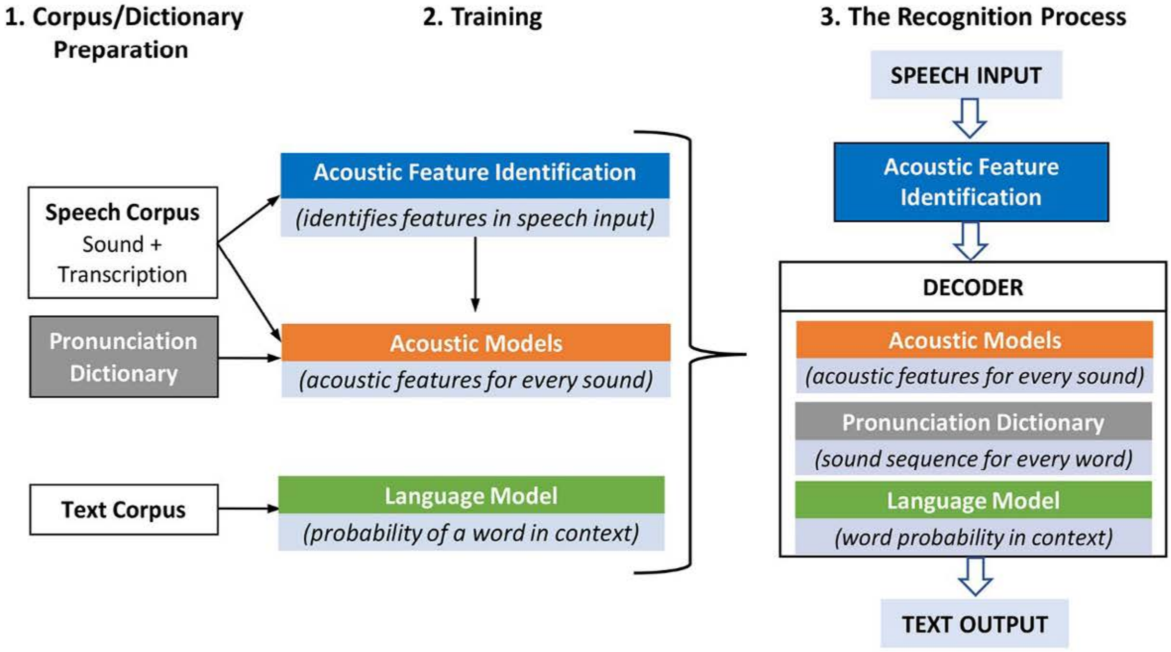

1. Corpus/Dictionary Preparation
2. Training
3. The Recognition Process
Speech Corpus
Sound + Transcription
Pronunciation Dictionary
Text Corpus
Acoustic Feature Identification
(identifies features in speech input)
Acoustic Models
(acoustic features for every sound)
Language Model
(probability of a word in context)
SPEECH INPUT
Acoustic Feature Identification
DECODER
Acoustic Models
(acoustic features for every sound)
Pronunciation Dictionary
(sound sequence for every word)
Language Model
(word probability in context)
TEXT OUTPUT

Figure 9.1: The Components of a Modular Speech Recognition System

2 **Training from corpora/pronunciation dictionaries**
The corpora and pronunciation dictionaries serve as training data for statistical, machine-learning techniques, used to build different parts of the recognition system. The acoustic (speech) signal and the corresponding sound transcriptions are used to train ***Acoustic Models***[1] i.e. characterisations of the acoustic features corresponding to each sound (vowel or consonant). As every sound is heavily influenced by the immediately preceding and following sounds, acoustic models are constructed for every possible permutation and combination of preceding and following sounds (see Chapter 2.4). Extensive text corpora (including the texts accompanying the spoken corpora) are used to build a ***Language Model***. *N-Gram Models* have been widely used: these allow prediction of the probability of each word, as a function of the preceding words. They allow the recogniser, when confronted with more than one likely word candidate to choose from, to pick the most likely candidate, given the preceding words. Usually, 2 or 3 preceding words are used for prediction. In recent years alternative language models, based on deep neural nets have increasingly been employed and allow word prediction based on both preceding and following word contexts.

3 **Decoding: recognising speech**
The acoustic features of the utterance are extracted from the incoming speech signal and fed to the decoder. The decoder calculates a probability for each possible sequence of sounds and words that might match the input: the probability is calculated on the basis of the acoustic model (for sounds in contexts) and of the language model (for words). The pronunciation dictionary connects the sounds in the acoustic model to the words in the language model. For both sound and word sequences, a number of possible candidates are retained, ordered from the most likely (best match) to the next most likely, and so on. The most probable sequence is output, although in some systems more than one option is retained.

9.3.2 End-to-End systems (E2E)

Advances in Deep Neural Networks (DNN) [3] and computational power in recent years are yielding increasingly accurate speech recognition End-to-End (E2E) systems [5-7]. As mentioned, these require vast speech corpus resources: when such resources are available for all the dialects, we can expect to see powerful *End-to-End* systems for Irish. These systems carry out the recognition process in a composite fashion, without relying directly on the individual components mentioned above [8-10]. Rather they deploy deep neural nets to find the correspondences between speech waveforms and text. Note that many E2E systems do, indirectly, use resources such as pronunciation dictionaries and language models. Also, note that even if not deployed in the recognition process, these resources are important in many applications that deploy speech recognition.

There are many challenges for Irish, if we aspire to applications that can meet the needs of the Irish language community. As emphasised in Chapter 1, technology development for a minority language like Irish is not driven by the commercial forces that hold in the major languages but must address the local (socio)linguistic factors that pertain and target those applications most critical for the language community and for language revitalisation. This is discussed further in §9.5 below.

9.4 Work done to date

Speech recognition systems are under development in the ABAIR group at Trinity College Dublin, and a modular beta-system ÉIST, can be accessed on its website[2]. Recently, an Irish ASR system has also been developed by Microsoft as part of its multilingual facility. The initial ÉIST system used as a starting point the phonetically annotated speech corpora developed for ABAIR's speech synthesis, supplemented by recordings of different dialects, obtained through a crowdsourcing platform *MíleGlór*[3] and by other available speech corpora. Other synthesis-related resources were also used, such as the pronunciation dictionaries for the main dialects, letter-to-sound rules etc. as detailed in Chapter 8. The focus was very much on catering for dialect diversity, and an approach to pronunciation building, drawing on earlier proposals in dialectology – for a 'common core' [11] to the Irish sound system – proved fruitful in constructing pronunciation dictionaries. The language models used in the system were developed in cooperation with Fiontar at Dublin City University, drawing on their extensive text resources, described in Chapter 3. For a description of ÉIST, see [12].

In parallel, End-to-End systems are being built, but

[1] Note that the machine-learned models (acoustic models and language models) discussed here are very different to the knowledge-based models described in Chapter 7 (see explanation of differences in Chapter 1.7).
[2] The beta speech recognition system, ÉIST, is available at https://www.abair/Eist. ABAIR is supported by the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media.
[3] Recordings can be made on the ABAIR MíleGlór platform at https://www.abair/mileglor.

they do not (yet) yield the accuracy of the modular system. These require vast corpora, and current research is focusing on corpus preparation and recording, particularly in Gaeltacht communities, to extend provision of state-of-the-art systems for speech recognition and synthesis for all dialects. Publicly available speech corpora are also being harvested: where parallel text is not available, transcriptions are being obtained in a semi-supervised way. The use of pre-trained self-supervised models is also being explored. These have the potential to achieve better performance with limited transcribed speech data by continuing this pre-training or by finetuning the model for the Irish speech recognition task.

A word-spotting application, currently being development by ABAIR—TG4, is aimed at enabling voice control of the TG4-player. Current research is also exploring the embedding of speech recognition into applications that already incorporate speech synthesis (Chapter 8), particularly applications for education and access.

9.5 Future work

There are specific challenges to providing recognition systems for Irish that are adequate for the diversity of the potential users and contexts of use, so that they can be truly useful for the language community. Whereas in the major languages, ASR systems initially targeted the standard variety, with later extension to further dialects, for Irish, all dialects need to be catered for from the outset (a system with an elevated failure rate for a particular dialect would be unacceptable). ASR for Irish must also recognise the speech of fluent non-native speakers. Furthermore, as educational applications are a priority for Irish, systems will be needed that can recognise the speech of learners with different levels of proficiency. Research on other languages has shown that recognition rates for existing systems are much poorer for learners, compared to native speakers [13].

Given the importance of educational applications, recognition of children's speech is a further priority. As explained in Chapter 2.4.2, the acoustic characteristics of sounds depend on the size of the speaker's vocal apparatus, something which changes continuously as children grow year by year. Even for those languages where speech recognition is most advanced, recognition of children's voices has tended to be the least successful.

The linguistic characteristics of the language can impact on the basic requirements for building a robust recognition system [14]. The large sound inventory of Irish (Chapter 2.4.2), and the fact that it is a highly inflected language (i.e. words occur in many forms, depending on the grammar: Chapter 2.5.1) have implications e.g. for the size of the corpora needed and the way in which pronunciation dictionaries are built, and these issues are discussed in [12]. The prevalence of code switching in Irish (frequent use of English words and phrases [15, 16]) presents further potential challenges for many future applications.

There are other factors – not specific to Irish – that are important in building recognition systems, e.g. the need for a system that doesn't fail when there is background noise as in a classroom, or when the speaker's style of speech switches from careful to rapid, casual speech.

Looking to the future, a different challenge stems from the fact that speech technology is moving towards dialogue-based conversational systems, where we interact with machines such as personal assistants or characters in games, in ways that simulate human-human interaction. As explained in Chapters 2.4.3 and 7, much of the meaning in our conversations arises not from the *words* we say (which speech recognition systems may recognise) but from the *way* they say them to express our attitudes and emotions (carried by variation in voice prosody, which speech recognition systems don't recognise). When communication is going well, the voice prosody of the speakers tends to converge: the extent of this convergence is indicator of successful communication [17]. The growing importance of this (unwritten) aspect of human spoken interaction is reflected in the growing field of *affective computing* and in the newly emerging applications, such as health monitoring, mentioned above. Ultimately, recognition systems will need to do much more than accurately track the words of the speaker.

In planning for Irish speech recognition, the big picture and the diverse requirements of future systems should be borne in mind, so that while the initial systems are being developed step by step, the foundations for the future needs are being accommodated. As with other technologies, the parallel development of key applications is vital to ensure that the benefits of research outputs are reaching the language community.

9.6 Recommendations

In the light of the above discussion, recommendations for developing speech recognition and applications over the coming years are to:

- **Leverage synergies with speech synthesis research:** a joint approach to research and development is recommended. The commonalities in corpus design and collection is discussed separately in 2 below. Beyond that, the two areas are converging and there are many areas where the pooling of knowledge, skills and resources accelerates progress in both domains. Both technologies need to evolve within a long-term framework, anticipating and developing joint strategies to meet the challenges discussed above, many of which are common to both synthesis and recognition, including multidialect coverage, children's voices, learners' needs, the requirements for dialogue systems with natural prosody, different styles of speech, code switching etc.

- **Very large-scale development of speech corpora:** ultimately, the quality and adequacy of the recognition systems will depend on the quality, the number and mix of speakers etc. in the speech corpora on which they are built. Note that corpus requirements of synthesis and recognition differ in significant ways (e.g. size of speech corpora, numbers of speakers, recording conditions). Nonetheless, there are many common challenges and a joint approach to the design, recording and processing of corpora will benefit both areas. A long-term strategy is recommended where ***corpus collection*** is carried out in close collaboration with the language communities and takes account of the various dialects, speaker age profiles, styles of speech, recording environments etc. For new recordings, the ***design*** of corpus materials needs to ensure, not only coverage of all sound combinations and contexts, but also that recording materials are suited to the dialect, to the speaker's age and to the envisaged applications – e.g. casual conversation as compared to formally read speech. In addition to new recordings, there is a wealth of publicly available speech corpora to be exploited. Where transcriptions are missing or incomplete, unsupervised or semi-supervised approaches may be used to train the ASR systems.

- **Continuous building, testing and evaluation of speech recognition systems as corpus resources emerge:** This should encompass the enrichment of current modular systems, as ***pronunciation dictionaries and letter-to-sound rules*** are extended for all dialects and ***acoustic models*** (from speech corpora) and ***language models*** (from text corpora) are improved. It should also include parallel development of End-to-End systems where corpus size is of great importance, and the use of pre-trained models.

- **Application-oriented constrained systems:** The ultimate Irish ASR system will recognise continuous speech of varied speakers (children, men, women with different dialects, learners etc.) including rapid informal speech, even in noisy environments. However, in the short term, it makes sense to build more limited systems targeting specific tasks. Many real-life applications are built for a specific context, and this allows restriction of the domain of recognition to the context of use, and to the envisaged end-users. This permits different strategies for optimisation of the system. There are different ways in which the speech recognition system can be constrained to targeted applications:

  Limited vocabulary, limited task systems are useful in many cases, e.g. educational applications where the entire vocabulary is predetermined. An application for young children could be optimised for children's voices.

  Speaker-dependent systems: in many applications, such as dictation systems, speaker-dependent systems are initially tuned to an individual voice, by having the user read certain materials the first time they use the system. This allows the recogniser to adapt to the speaker's acoustic models. Dialect-adaptation of recognition systems is also likely to be helpful.

  Isolated word/phrase recognition systems: applications where the speaker pauses between words/phrases can greatly improve recognition performance: even a simple dictation system with isolated word recognition has many applications, e.g. composing essays, emails, books, and is useful for those with dyslexia or visual impairment etc.

  Word spotting: an application that uses ASR at the front end and carries out some further action, such as controlling a device. Such applications make more limited demands of the speech recognition system, requiring only a limited number of words to be spotted, but by speakers of different ages, dialects etc., as in the application to control the TG4 player, mentioned above.

The choice of application type to focus on should be determined by its potential usefulness in the Irish context, as well as by what may be feasible at a given time with the resources available. Ongoing interaction with the Irish language community is therefore essential to inform application priorities. And although systems that can deal with the diversity of end user needs (ages, dialects, speech style etc.) may take time to deliver, every advance towards that goal will enable interesting and useful applications, as well as enabling fledgling dialogue systems and creative new vistas in education, disability/access and public use.

References

[1] Rabiner, J. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

[2] Chiu, C.-C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E. & Jaitly, N. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4774-4778).

[3] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M. & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access 7*, 19143-19165.

[4] Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: Timit and beyond. *Speech Communication 9*(4), 351-356.

[5] Jurafsky, D. and Martin, J. H. (2019) *Speech and Language Processing*. Prentice Hall.

[6] Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation 18*(7), 1527-1554.

[7] Mohamed, A.-r., Dahl, G. & Hinton, G. (2009). Deep belief networks for phone recognition. In *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications.*

[8] Yu, D. & Deng, L. (2010). Deep learning and its applications to signal and information processing. *IEEE Signal Processing Magazine 28*(1), 145-154.

[9] Chorowski, I.K., Bahdanau, D. Serdyuk, D., Cho, K. & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems 28*, 577-585.

[10] Chan, W., Jaitly, N., Le, Q. & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4960-4964).

[11] Ó Murchú, M. (1969). Common core and underlying forms. *Ériu 21*, 42-75.

[12] Lonergan, L., Qian, M., Ní Chiaráin, N., Gobl, C. & Ní Chasaide, A. (2022). Cross-dialect lexicon optimisation for an endangered language ASR system: the case of Irish. In *Proceedings of Interspeech 2022*.

[13] Tomokiyo, L.M. (2001). *Recognizing non-native speech: Characterizing and adapting to non-native usage in LVCSR*. [Ph.D. thesis]. Carnegie Mellon University.

[14] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication 56*, 85-100.

[15] Ní Laoire, S. (2019) *Códmhalartú, códmheascadh agus trasteangú: teangacha i dteagmháil [Codeswitching, codemixing and translanguaging: languages in contact*]. In T. Ó hIfearnáin (Ed.) *An tSochtheangeolaíocht: Taighde agus Gníomh* [*Sociolinguistics: Research and Practice*]. Dublin: Cois Life, 49-78.

[16] Ní Laoire, S. (2016). Irish-English code-switching: a sociolinguistic perspective. In R. Hickey (Ed.) *Sociolinguistics in Ireland*. London: Palgrave Macmillan, 81-106.

[17] Weise, A., Levitan, S. I., Hirschberg, J., & Levitan, R. (2019). Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication 115*, 78-87.

Chapter 10
Spoken Dialogue Systems

10.1 What are they?

Spoken dialogue systems (SDS) enable people and machines to interact using natural language. Instead of using technical menus or complicated computer coding, people can simply speak to or text the SDS on their mobile phone, tablet or laptop, and get responses in either natural speech or text. These systems are complex pieces of software that draw on a number of artificial intelligence (AI) technologies. The building blocks of natural language or dialogue interfaces, speech synthesis (TTS), speech recognition (ASR) and natural language processing (NLP), are well established for English, and are either available or in development for Irish, as described in Chapters 4, 8 and 9 of this Digital Plan.

Early dialogue systems were text-based, but recent advances in spoken language technology have made speech between user and machine feasible. Phone-based virtual assistants, like *Siri*, are examples of modern speech interfaces, and more sophisticated systems are widely used in the areas of banking, healthcare, retail/customer service and education. Natural language interfaces, and particularly voice, are considered major growth areas in technology for the next decades. Spoken interaction is the medium for most human activity, and so, voice technology allows computers to perform in situations where speech is the most natural and often the only feasible interface, particularly in disability/access contexts, infotainment, edutainment and serious games.

SDS are inherently cross-platform and multimodal – systems can communicate using speech, text or both. The interface can be as simple as a mobile phone messaging application (Facebook Messenger, WhatsApp), or as complex as a fully featured human-like on-screen avatar, or even a physical robot.

10.2 Why are they important and for whom?
10.2.1 Current systems

Table 10.1 gives examples of where dialogue systems are currently used. The most frequent use tends to be commercial, e.g. call centres. Many of the examples would be rather low priority in the case of Irish and the shaded grey boxes of the table show those that would be of most interest. They are elaborated on below.

Examples	For Whom?
In-car navigation	General public
Call centres	General public: interface to public (business & institutions)
Automatic telephone help desks	General public: interface to public (business & institutions)
Health interfaces	General public: e.g. providing health information – see Figure 10.1, *SimSensei* a virtual human dialogue system that conducts interviews related to psychological distress conditions, such as depression, anxiety, and post-traumatic stress disorder
Social interfaces and dialogue-based instruction manuals	General public: e.g. interactive talking manuals to acquire new skills, troubleshoot,etc
Siri, Alexa, Cortana, Google Assistant, etc.	General public (personal assistant) / Disability & access
Intelligent tutoring systems	General public/Learner populations
Conversational agents	General public/Learner populations
Pedagogical agents	Learner populations
Immersive information-giving applications	General public/Specific cohorts: e.g. museum guides/classroom aids (Figs 10.2, 10.3)

Table 10.1: Examples of where Dialogue Systems are used. This illustrates a current health interface which allows users with psychological disorders to talk to a virtual healthcare specialist.

Figure 10.1: SimSensei & MultiSense: Virtual Human and Multimodal Perception for Healthcare Support (https://www.youtube.com/watch?v=ejczMs6b1Q4)

Question-answer systems, e.g. phone-based virtual assistants: as with Siri on mobile phones, one asks a question, the system searches a database and provides a spoken answer. It is not available for Irish but as systems like Siri are widely available for English they are increasingly being requested for Irish, especially by people with visual impairment. Note that these systems for English have access to enormous databases, which are not available for Irish. Furthermore, speech recognition for Irish is at an early stage of development.

Intelligent tutoring systems (ITS): using spoken language technology and natural language processing, systems can be built that offer personalised tuition to learners in many domains. Such intelligent tutoring systems assist learners by posing questions, parsing responses and offering customised instruction and immediate feedback, usually without needing teachers' input. When designed carefully, these systems are enormously useful in complementing classroom-based instruction, allowing remote users access to material, and providing extensive extra practice at low or zero cost. Such systems have been successfully deployed, mainly in the areas of physical sciences and mathematics, and are now appearing in health sciences, language acquisition and other areas of formalised learning.

Modern ITS systems provide instant feedback at the session level, and allow detailed continuous assessment by maintaining and updating a learner's profile. These systems both record and predict student abilities and future needs, thus facilitating the provision of a tailored and dynamic curriculum. The system computes a statistical estimate of the student's mastery of the concepts and skills that underlie an exercise. It will continue to present a student with problems that test a concept until the application's logic determines that the student has mastered the underlying knowledge as it continues to gather analytical data on all those who interact with the system.

Until very recently, such classic ITS systems have been impersonal, instructivist, non-social and abstract pedagogical agents whose sole purpose is to tutor. These systems tend to lack any personal characteristics and therefore students do not develop personal connection with them, removing intrinsic motivation.

Pedagogical agent: this refers to a computer-generated character employed in an educational setting in order to fulfil specific pedagogical tasks [1]. The concept of an 'agent' denotes an entity with some degree of 'intelligence' and capacity for autonomous action. In contrast to classical tutoring systems, it is generally assumed that pedagogical agents are visually embodied and often animated.

McTear wrote in 2002 that 'the "conversational computer" has been the goal of researchers in speech technology and artificial intelligence (AI) for more than 30 years' [2] and this goal, long elusive, is being very actively pursued today. Essentially conversational agents allow the user to engage in a two-way conversation with a virtual computer-based 'listener' – i.e. agents that interact with people through language to assist, enable or entertain.

Conversational agents: human conversation is very different from having text read aloud. As discussed in Chapters 2, 7 and 8, the interactive expressive nature of human-human conversation is not currently available for dialogue systems, but this is where the technology is heading. Conversational speech is key to building a strong working relationship between system and user - the inclusion of conversational interaction has been shown to increase engagement, retention and adherence in dialogue applications in education and healthcare. In such systems, the human-like interaction causes the user to view the system as a partner rather than a simple machine. This phenomenon is of immense value to educational applications.

The term 'spoken dialogue system' is used for a wide range of facilities. Some are very restricted in terms of the topics and often involve a single domain such as flight enquiries. Some might even be more restricted, e.g. allowing only the digits 0–9 and the words yes and no, while others permit large vocabulary systems and relatively freeform 'conversational' input. The ultimate aim, however, is for true 'conversational' systems that are capable of engaging in an extended conversation with the user on any topic.

Immersive information-giving applications: one example is shown below in Figure 10.2, and is a virtual reality museum guide. Visitors can have one-to-one, face-to-face natural language interactions with the guide, ask questions and get answers. Being able to engage people young and old by delivering information in an interactive setting would have many benefits as a way of widening the domains of usage for Irish speakers and Irish learners. It would be relatively straightforward to realise such an application in Irish museums, galleries and interpretive centres as Irish language descriptions of the exhibits are already available. The fact that these settings provide a limited-domain, make them an ideal testbed for nascent dialogue systems.

Experimental work in the field of education is ongoing where holograms and bots are used to preserve the stories of people who have lived through various experiences (see, for example, Figure 10.3, a conversation between a real Holocaust survivor rendered as a virtual personage). This could have a much more powerful impact on learning about history and language compared, for example, to reading second- or third-hand accounts of that experience, as it enables a conversation between today's generation and a preserved model of the older generation. As with many emerging dialogue systems, collaboration with many areas, including virtual reality and multimedia research, will be needed.

Figure 10.2: Museum of Science Boston's Virtual Human Museum Guides (https://www.youtube.com/watch?v=rYF68t4O_Xw)

Figure 10.3: Natural language interaction via hologram: From The Near Future - David Traum, USC Institute for Creative Technologies (https://www.youtube.com/watch?v=UjI-W1z9ruw)

10.2.2 Aspirations for Irish

While education has been central to many of the applications developed to date, conversational dialogue systems are increasingly becoming part of the everyday environment and we aspire to systems for Irish that will be on a par with those emerging for the major languages, and which can be used in a multitude of environments.

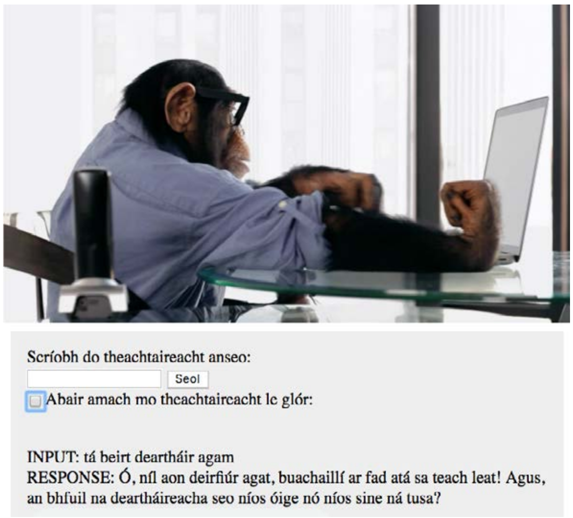
In education, the gold standard aspired to for Irish will be a ***hybrid intelligent conversational pedagogical agent*** – an agent that is capable of engaging in a conversation with a learner. Ideally, one could combine the ***pedagogical*** function of an ***intelligent tutoring*** system with natural language ***conversation***. The conversational tutor would ideally be situated in an immersive environment that would engage the learner and give them a sense of having a real/meaningful interaction with the virtual interlocutor, so that a ***personal connection*** is formed.

10.3 What has been done to date for Irish?

This is a very new field for Irish. : However, a fledgling dialogue system, *Taidhgín*, has been developed for Irish at the Phonetics and Speech Lab, Trinity College Dublin [3, 4].

Taidhgín* as a conversational partner** takes the form of an animated monkey which can discuss a number of topics with the user. *Taidhgín* speaks with an ABAIR synthetic voice and the user chooses from the selection of other available voices which one they wish to represent themselves. *Taidhgín* at this point is an initial prototype: the goal is that it will be used as an ***intelligent language tutoring system, a ***pedagogical*** agent and a ***conversational*** agent, i.e. incorporating as far as possible the kinds of features envisaged for future development.

Given that speech recognition was not available for Irish at development stage, the user inputs text, which is then spoken aloud. Taidhgín responds with speech output but the text is also presented to the learner. Early evaluations of the prototype have shown that the system can be successfully used to engage learners in a conversational interaction using a natural language interface. The user feedback showed that despite the limitations of the current system, learners were quickly engaged and enjoyed their conversations with the 'intelligent' monkey.



Scríobh do theachtaireacht anseo:

□ Abair amach mo thcachtaircacht lc glór:

INPUT: tá beirt deartháir agam
RESPONSE: Ó, níl aon deirfiúr agat, buachaillí ar fad atá sa teach leat! Agus, an bhfuil na deartháireacha seo níos óige nó níos sine ná tusa?

Figure 10.4: Taidhgín, a social conversational pedagogical agent, shown in mid-conversation.
Note that all conversational turns are logged for subsequent analysis.

***Taidhgín* as an intelligent Tutor:** more limited versions are also being developed where *Taidhgín* serves as an intelligent tutor, who trains learners on targeted spelling and grammatical features of the language, along with training on non-language-specific content. These specific applications, *Trialacha Taidhgín*, are presented in the form of mini-quizzes/games. These will focus more on integrating existing linguistic knowledge as well as available and newly emerging language technology into the games.

Taidhgín, and other such animated agents, are envisaged as central to future CALL development (see Chapter 14). Although currently at a very embryonic stage, it is hoped that interactive agents of this kind will ultimately be able to:

- engage in social interaction – e.g. being able to converse with learners using natural language in order to create a positive and enjoyable learning experience
- provide intelligent tutoring in specific areas (e.g. pronunciation, grammar, vocabulary etc. or for teaching subjects through the medium of Irish, e.g. science)
- adapt to individual learners' needs: be able to offer help on request or recognise the kinds of recurring mistakes that a learner makes and provide relevant personalised feedback

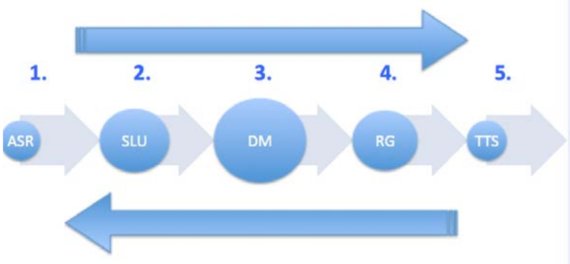


Figure 10.5: Basic spoken dialogue system architecture

10.4 How does it work?

Figure 10.5 illustrates the most basic steps involved in the operation of a spoken dialogue system.

Spoken dialogue systems are complex to set up since implementation requires employing a number of technologies to process human language, each of which are distinct fields in their own right (many are already described in other chapters in this Digital Plan). The five steps shown in Figure 10.5 are listed here and elaborated below:

(1) **ASR:** *(automatic speech recognition)* of speech input
(2) **SLU:** *(spoken language understanding)* extracts aspects of meaning from the speech
(3) **DM:** *(dialogue management)* is the central, controlling component of a spoken dialogue system – it tries to decide what is the best response ('action')
(4) **RG:** *(response generation)* generates natural language response
(5) **TTS:** *(text-to-speech synthesis)* converts response to speech for output

The user talks into the system and it converts the speech to text (1). Aspects of the meaning are then extracted (2). Note that this could be simple word spotting (as in the current version of *Taidhgín*) or a more sophisticated analysis. The dialogue management is the heart of the system and allows the identified meaning to be matched (in a simple system) or interpreted (in a more complex system). This, in turn, determines the *action* required (3), i.e. the response that is likely to match, which will also serve to prompt the user for further input leading to another cycle through the process. The next step (4) is to adapt this response so that it is appropriate to the dialogue context to date (e.g. follows the tense, etc.). The final step converts this text response to a speech output (5).

Dialogue systems bring together many different core technologies and linguistic resources. Some of these are already available, such as Step (1) speech synthesis (Chapter 8). The research to provide expressive voices for conversational interaction (Chapter 7) promises to bring us closer to a true conversational experience. Step (5), speech recognition (Chapter 9), is also important but is still in the early stages of development. As discussed there, targeted recognition systems that can provide for specific educational contexts would make *Taidhgín* more powerful as a pedagogical tool. Steps (2) and (4) above will become available through the NLP research described in Chapters 4 and 5 and NLG in Chapter 6. The vital core of the system (Step (3)) is the dialogue manager. This is the central research activity that will be required, along with integration of the other components above.

Rule-based vs. machine-learning approaches

As with many other technologies, there are two possible ways of going about the development of the dialogue system (see discussion in the Chapter 1.7). A rule-based system will be needed initially that draws on linguistic knowledge, understanding of the real-world context, models of speech prosody for dialogue interaction, annotated dialogue corpora, etc. At a certain point, where sufficiently large corpora become available, machine-learning strategies can be introduced into the process. However, caution will be needed: simple adopting a machine-learning system based on English is unlikely to be a fruitful avenue for Irish (see Chapter 1).

10.5 Recommendations: a phased approach

The development of dialogue systems will need careful planning and scoping. It is important that the systems be developed with specific high-priority applications in mind – it is envisaged that educational and disability/access applications will be central for Irish. The process of providing dialogue systems will entail the following broad steps:

- **Irish language dialogue manager:** build a system that will allow for the development of sophisticated versions of platforms such as the current Taidhgín prototype, exploring the use of different kinds of dialogue partners, including:
  - intelligent tutor
  - pedagogical agent
  - conversational agent
  - interactive/immersive platform
- **Research on conversational expressive interactive speech, entailing, for example:**
  - gathering and processing of spoken dialogue corpora. In principle this should be carried out with the collection of spoken corpora intended for synthesis (Chapter 8) and recognition (Chapter 9).
  - research on expressive speech, in collaboration with the work of Chapters 2, 4, and 7. This would entail ongoing research on speakers' productions and listeners' perceptions of the critical expressive and conversational features.
- **Integrated system development:** where the various speech and NLP resources can work together in a way that allows streamlined development of specific real-life applications.
- **Content development:** content and learning scenarios will require close collaboration with subject matter experts.
- **Develop external collaborations with multimodal and virtual reality groups:** joint projects that will allow for the extension of platforms to include, for example, agents with a variety of personalities, visual representations, virtual reality settings.
- **A programme to develop real applications that harness the emerging technologies and resources:** Provision will be needed to allow prototype systems to move from fledgling to stable products that can be released to the public.
- **Ongoing evaluation:** targeted user groups will be integral to every stage of development, including the initial design stages.

Overall, an iterative loop of development is envisaged that will involve using the knowledge-bases and resources of Part I as well as the core technologies of Part II. As dialogue systems come on stream, they will be central to many of the applications of part III, and especially to CALL (Chapter 14).

References

[1] Gulz, A., Haake, M., Silvervarg, A., Sjödén, B. & Veletsianos, G. (2011). Building a Social Conversational Pedagogical Agent: Design Challenges and Methodological approaches. In D. Perez-Marin & I. Pascual-Nieto (Eds), *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*, pp. 128-155.

[2] McTear, M. (2002). Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing Surveys 34*(1), 90-169.

[3] Ní Chiaráin, N. & Ní Chasaide, A. (2016a). Chatbot Technology with Synthetic Voices in the Acquisition of an Endangered Language: Motivation, Development and Evaluation of a Platform for Irish. In *Proceedings of the 10th Language Resources and Evaluation Conference* (pp. 3429–3435).

[4] Ní Chiaráin, N. & Ní Chasaide, A. (2016b). Faking Intelligent CALL: the Irish context and the road ahead. In Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, (pp. 60–65).Björkenstam, &amp; L. Borin (Eds.),*Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC* (pp. 60–65). Umeå, Sweden: Linköping University Electronic Press, Linköpings universitet.

Chapter 11
Machine Translation

11.1 What is machine translation?

Automatic Machine Translation is a software-driven process transforming digital content (text, speech, etc.) from one language to another. This is a very complex task which involves taking the meaning and conventions of a source message and generating a message conveying the same meaning using the conventions of the target language.

The technology behind modern machine translation (MT) systems is constantly advancing. However, there are some dominant paradigms in the field, which are namely data-driven (statistical and neural MT) and rule-based (linguistic-based MT). Both approaches have their strengths and weaknesses. These days, both in research and industry environments, neural MT is the most widely adopted paradigm.

The purpose of machine translation is two-fold. One purpose is to get the gist or sense of a piece of text in a language unknown to the user. The other is for pre-translation purposes in a professional environment, where the translation from a MT system is then proof-read or post-edited by a professional translator. The reader will be familiar with the inadvertent misuse of free online MT systems, whereby the professional post-editing step is often overlooked. This is largely due to the fact that it is often not made clear to the user that checking and proofreading of automatically translated output is essential. In addition, the quality of free online MT systems varies greatly from translation pair to translation pair, where no indication of the relative quality or reliability of one particular translation pair over another is given.

This chapter sets out to explain how MT works, how it can be improved for Irish language use, and how, when used as part of a translation process, it can serve to reduce translation costs and effort.

11.2 Why is it important and for whom?

The job of the translator has changed significantly over the past 20 years through the introduction of translation technology. In order to help reduce translator workloads, improve productivity and help narrow the gap between supply and demand, the translation and language services industry is striving to provide translators with improved tools with which to carry out their work. Historically this has taken the form of glossaries, terminology databases and Translation Memory (TM) software. More recently as the underlying science, software, and infrastructure to support it has improved, the use of Machine Translation (MT) tools has also become standard practice in professional translation environments across the world. The use of MT in public administration has been widely adopted in bilingual countries, such as Canada[1] and Wales[2].

Modern MT tools can be used in a variety of ways. Typically they are deployed within a standard translator's workbench as an additional support tool. In this way the translator carries out their work as normal, using glossaries, term banks or TM software, but with the added option of receiving suggestions generated automatically by MT as well. This type of seamless integration allows the translator to use what is most useful and effective for the task at hand and they can choose to work from scratch, to combine elements from the various supports available or to simply post-edit a useable suggestion generated automatically by the MT engine.

For many years, extensive studies have shown how the integration of MT within such a workflow (often complementary to the use of translation memory tools) improves productivity, both in industry-based and in academic-based research, e.g. [1, 2]. Additionally, it has been shown that when working on new or unfamiliar texts, where translation memory matches would be low, translators have higher productivity and quality when using machine-translated output than when translating without it. The results of these studies point not only to a significantly increased capacity for translators using MT

[1] https://nrc.canada.ca/sites/default/files/2019-03/dt_evaluation_report_2018_e.pdf
[2] http://www.assembly.wales/en/abthome/about_us-commission_assembly_administration/Pages/machinetranslation.aspx

tools, thereby addressing the supply/demand gap, but also to significant cost savings which can be achieved. This in turn means that a certain amount of the additional work needed to meet (increasing) demand can be achieved at no additional labour cost through the use of more modern translation tools that utilise MT.

Additionally, and of significance for a minority language like Irish, the status of the language can be bolstered through improved translation services that will result in an increase of Irish digital content, thus protecting citizens' rights to interact with the state and services through Irish. At a national level, there is a strong demand for English to Irish (EN>GA) translation systems within the Irish public services and other businesses operating in Ireland and serving the Irish market. In particular, with respect to legislative translations, a backlog has arisen whereby the demand for translation outweighs the supply [3]. Compared to the volume of content that needs to be translated, there is a relatively low number of suitably qualified translators with a relatively limited output capacity. Currently, the gap caused by this situation is widening, since the lifting of the derogation on translation of EU legislation and documents at the end of 2021. As with all European languages, translation technology will play a significant role in closing this gap both nationally and at a European level.

11.3 How does machine translation work?

The design of many MT systems is what we term, 'data-driven'. Computer engineers feed (train) the system with parallel data (previously professionally translated text) from which it learns how to predict a new translation (see Figure 11.1 for how Statistical MT works). A statistical algorithm works out correspondences between strings of characters in each language. Then, based on how often it sees these correspondences, the algorithm assigns a probability, which is used to determine the most likely translation for a given input string. The more examples of translations it sees, the better the predictions and the higher the accuracy. The generated translation is then improved with a language model, trained on monolingual text of the target language. Neural machine translation, on the other hand, is based on machine learning and neural networks [4-6]. It involves taking parallel training data and converting the words into vector representations (mathematical encodings), before passing a representation for the entire sentence through a neural network. This way, the context of a word is much better understood and accounted for. NMT systems are known as 'end-to-end' systems as they do not have the separate components that SMT systems have.

These data-driven approaches have advantages in that they are quick to implement, quite robust when dealing with bad/noisy data and can be tuned to translate text within a specific domain (e.g. medical, legal, public administration). However, both types of approach rely on the assumption of an availability of suitable training data and in sufficient quantity (NMT requiring more than SMT), which, in the case of Irish, is often hard to find.

Sourcing good parallel data can be a challenge, particularly in the case of lesser-resourced languages such as Irish. Often the main task of creating a parallel corpus involves the identification and collection of previously translated data from translators or translation bodies, and contacting appropriate government departments or semi-state organisations for their legacy translations. In addition, software tools called 'web-crawlers' are used to find parallel texts on the internet. Ensuring correct alignment

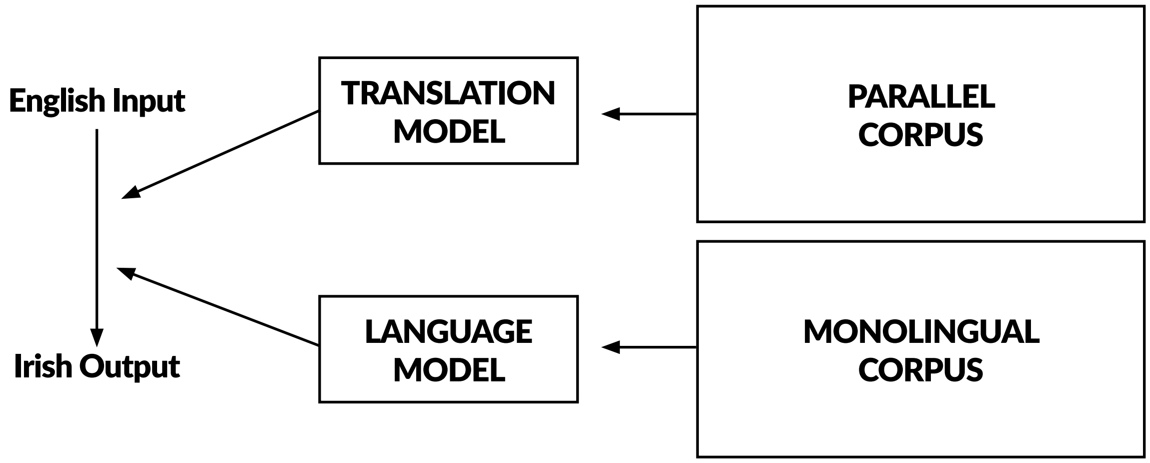


Figure 11.1: The architecture of a basic Statistical Machine Translation system

across sentences to prevent the MT system learning incorrect translations is vitally important. It should also be remembered that while SMT and NMT are often good at handling difficult idiomatic language, without a rule-based component it can also make basic grammatical and logical errors.

A rule-based MT system, on the other hand, uses the inherent linguistic rules and regularities of a language to create a translation. This method involves linguistic engineers encoding a computer with human knowledge from the grammar and linguistic structure of the source language and mapping it into the target language. This information is then used to generate the final output translation using knowledge about the grammar and linguistic structure of the target language. This approach is more elegant in many ways and can cater very well to nuances of language that are difficult to glean from data alone. However, it relies on the availability of tools for each language that can carry out the linguistic processing of data, as well as manual and highly specialised work to encode the various rules needed. This means that systems using rule-based MT take longer to implement and are less easily converted from one language pair to another.

For both paradigms, there are reference implementations, which are leaders in both research and industrial applications alike. Both types of toolkits are available through the open source platforms Apertium[3], Moses[4] and OpenNMT[5] for rule-based, statistical-based MT and neural MT respectively. Current ongoing work by Irish research groups using these platforms is yielding positive results in developing MT engines for EN>GA translation [7-9].

11.4 What has been done for the Irish Language?

In recent years there has been a number of studies on machine translation for Irish. On the statistical system side of things, researchers at DCU developed a Moses-based MT system for use by in-house translators at the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media (DTCAGSM) as part of their translation workflow [7]. The Tapadóir system can be used to pre-translate the text so that professional translators would only need to correct (post-edit) the system's suggested translation. The system reaches a quality bar that has been recognised as suitable for a professional post-editing setting, mainly due to being specifically trained on previous translations held by DTCAGSM.

Studies on various methods to improve Irish MT continue to be carried out in DCU through postgraduate studies. Machine-translation experts have also assessed the use of Neural Machine Translation for Irish, along with identifying ways to improve algorithms to achieve better results despite having limited training data to work with [10]. Linguistic modules are also being integrated into these systems to ensure that post-editing efforts required by translators are minimised [9]. Research on SMT and NMT for Irish was also carried out at the Insight Research Centre, NUI Galway [11-12].

Research into MT for Irish in Trinity College, Dublin has focused mainly on the linguistic aspects of translating between English and Irish. A number of studies have been carried out in recent years [13-16] investigating aspects of the use of rule-based MT systems such as Apertium [8] with rule-based parsing technology for Irish [17], to be used in hybrid MT solutions. Studies investigating the use of linguistic information in Neural Machine Translation systems show promising results and this is likely to be a fruitful avenue of research for the English-Irish pair.

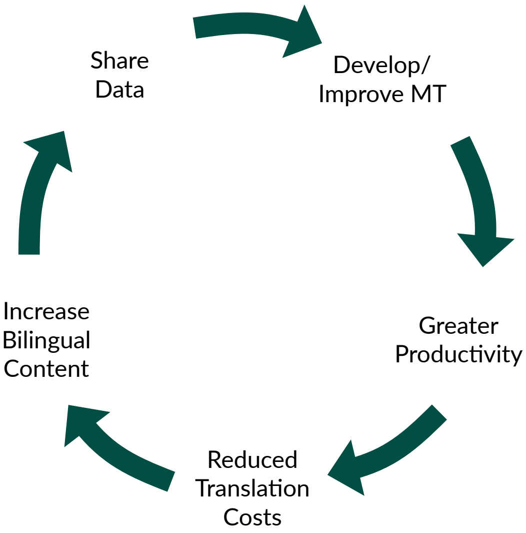


Figure 11.2: The virtuous cycle of sharing data to improve machine-translation systems

[3] http://www.apertium.org
[4] http://www.statmt.org/moses/
[5] https://opennmt.net/

As quality data-driven MT systems rely on the availability of quality language data (parallel or monolingual), it is worth noting the ongoing efforts in collecting and coordinating such data across the country. In May 2019, through the European Language Resource Infrastructure (ELRI) project, a national protal was launched in Ireland by researchers at DCU. The eSTÓR portal[6] (formerly NRS) facilitates Irish language data collection in which users can upload their own bilingual data and terminologies to the portal, and in the case of uploading batches of parallel files, translation memory files will be automatically created for the user. Anyone working with the Irish language within public administration can request an account for the portal. All of the data being collected will not only be used within the European Commission's eTranslation system (Neural MT), but can also be used by any Irish public administration body that would like to access translation memory files or that seeks to develop their own bespoke MT system. For example, the EU-funded PRINCIPLE project saw bespoke MT systems being developed and used by translators within Rannóg an Aistriúcháin, Foras na Gaeilge and National University of Ireland Galway. Figure 11.2 demonstrates the long-term benefits of data sharing for potential users.

11.5 Recommendations

Continued development of modern and advanced software is required to ensure the availability of a high-quality English-Irish machine translation infrastructure. The existing work done in this area has shown the viability and positive impact of this technology and approach, however, it is not yet comprehensive enough nor widely adopted in many sectors. In order to support continued advancement in this field, five significant lines of action have been identified as areas that require immediate attention. They will provide the necessary underpinnings to allow the existing translation community, as well as the academic and industrial R&D teams, to continue to make innovations in this area in order to provide the tools and resources needed to ensure that our language is appropriately supported by the most up-to-date translation tools and resources.

- **Data collection and sharing**

Whether for training statistical systems or for validating and testing linguistic processing components, digital parallel corpora are extremely valuable data sets for developing MT tools (see Chapter 3 on Corpora). Across the country, valuable language data is being underused and undervalued. This type of data is readily available as a by-product of the publication of content in Irish or of English-Irish translation work. All that is required, is for this data to be collected, stored appropriately and made available for reuse under a suitably permissive licence, (e.g. as CC-BY[7], which is the selected licence for the reuse of public data under the EU Open Data Directive)[8].

This data should be in a suitable machine-readable format like XML, TMX, XLIFF etc. (or any of their successor/derivative formats) which will maximise the ability to reuse and repurpose the data automatically without the need for laborious manual alignment. Where possible this data should be shared freely for the benefit of the community through portals such as the National Relay Station (eSTÓR) and a culture of data sharing should be encouraged amongst translation stakeholders, which will help foster a virtuous cycle of data provision and reuse. It is also important that data holders are educated on the value of their language data. Advice should be provided on how best to draw up translation contracts with third-party translation houses to ensure that, where possible, translation memory files, or any such by-products, are returned to the state-funded organisation or department that is financing the translation request. These TMs can then be used to negotiate fair translation costs going forward, as well as serving as training data for national MT systems.

Such data management practices will be supported by the Open Data Unit's policy[9] on making Ireland's public data digitally accessible and should be upheld across those working with Irish-language data within the relevant bodies.

- **Interface with wider European initiatives**

Where at all possible, further development in this field should be undertaken in conjunction with, or with an awareness of, wider initiatives in the field. There are existing community-driven and EU-driven efforts around collecting and sharing linguistic data and tools, e.g. META-SHARE, CLARIN, ELRC, ELRI and the European Language Grid (ELG), which can be used to the benefit of Irish.

[6] https://estor.ie
[7] https://creativecommons.org/licenses/by/4.0/
[8] https://data.gov.ie/pages/open-data-directive
[9] https://data.gov.ie/

Ongoing maintenance and support of the eSTÓR is necessary in order to facilitate better central management of national translation data, and to ensure reuse and leverage of existing resources where necessary. This portal should be connected to the Irish Open Data Portal and facilitate integration with any future centralised systems such as a Shared Translation Service.

At European level, this approach for training data-driven MT systems is being embraced (eTranslation)[10] and is already available for use by those in public administration across Europe. The English-Irish version of the eTranslation system is, however, still currently below par compared to most other EU languages. The main reason for this is the lack of sufficient English-Irish translation data available to train the system during the derogation.

The European Commission has proposed for the first time a series of guidelines covering the objectives and various digital service infrastructures including automatic translation – CEF.AT[11] – and the Digital Europe Programme[12]. At a national level, efforts were already made to increase the availability of training data for the eTranslation system through the Connecting Europe Facility (CEF) programme through the ELRI and PRINCIPLE projects. Ireland's involvement in the European Language Resource Coordination (ELRC) has also played a central role in the advances of Irish MT. The Irish administration overseeing the advancements of translation and translation technology will need to ensure alignment with such EU activities and programmes, as there is much benefit to be drawn from cooperation with other member states with respect to their advances in techniques and infrastructures.

- **Further research and development of MT technology for Irish**

The future of Irish MT cannot lie in the hands of global technology companies such as Google or Microsoft. In particular, professional translation requirements such as those within public administration should be carried out in secure environments such as eTranslation or bespoke MT systems that have been tuned specifically to the genre of text required. Progress in developing the Irish systems for eTranslation would benefit from ongoing collaboration with the EU's Directorate-General for Translation (DGT). Bespoke MT systems can be developed through supporting Irish universities or research groups.

The initial research and development noted above has already shown the potential for development of machine translation technologies for Irish. These initial short-term projects show the value of investing in projects and R&D programmes supporting MT development. In addition, this research should be broadened to encompass the emerging field of speech-speech translation, which combines the disciplines of speech technology and machine translation.

Further support for long-term programmes and projects is paramount in order for these advances to be realised and for the properly researched and developed technologies to be useable.

- **Human evaluation of machine translation output for Irish**

Automatic metrics (e.g. BLEU [20] and Meteor [21]) are often used as guidelines to indicate the quality or reliability of the translation output of a machine translation system. However, these metrics are not a reliable indicator of how useful this output is in a professional post-editing environment. Human evaluation involves asking experts in the target language to rate a translation and also to measure the amount of effort required to correct MT output. The ability to objectively assess translation technology tools according to their post-editing effort is essential for ensuring a well-informed decision as to which translation tools to use. It also ensures that tools produce translations that require less effort to post-edit than the effort that would be necessary to translate the same texts from scratch [22]. Not only does the translation quality of a MT system need to be measured, but also the effectiveness or usefulness of MT systems to translators in specific translation environments [23].

While some research has been carried out to date on human evaluation of Irish machine translation (comparing DCU's SMT and NMT engines with Google Translate), as yet there is still no clear picture on how the quality of MT impacts the job of an Irish translator. This is mainly due to (a) the limited use of Irish MT in professional environments to date and

[10] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[11] https://ec.europa.eu/digital-single-market/en/automated-translation
[12] https://digital-strategy.ec.europa.eu/en/activities/digital-programme

(b) the current lack of evaluators with the necessary combined skillset.

Types of human evaluation activities include [24]:

i) Assessment of fluency/adequacy of each translated sentence on a Likert scale

ii) Ranking assessment of either:

– different MT systems (e.g. free online system and a domain-tailored system) or

– different implementations of a domain-tailored system based on varying degrees in training data sizes or types

iii) Measuring post-editing effort: temporal (time taken to post-edit) [25], technical (number of edits required) and cognitive (how much effort is required to post-edit, measured with an eye tracker) [26]

- **Education and training**

There are a number of areas in which investment in focused MT training and education is required:

- Machine translation technologists ideally should possess the combined skills of computer engineering, linguistic theory and Irish language. The combination of skills ensures that the engineers or developers of MT systems are sufficiently equipped to analyse the output of a machine translation system in development, and thus tune it appropriately.
- All third level Irish translation studies and related courses should include modules that ensure that graduates are sufficiently trained to use current translation technologies (including translation memory software and machine translation tools). This includes education on the level of responsible data management, data sharing and the importance of TM reuse.
- Post-editing Irish machine translation output is a specific skill that all professional translators should be familiar with [27].

References

[1] Bywood, L., Etchegoyhen, T., Georgakopoulou, P., Fishel, M., Jiang, J. Loenhout, G., Pozo, A., Turner, A., Volk, M. & Maucec, M. (2014). Machine translation for subtitling: a large-scale evaluation. In *Proceedings of the 19th International Conference on Language Resources and Evaluation* (pp. 46-53).

[2] Guerberof Arenas, A., (2008). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus* 7(1), 11-21.

[3] Walsh, J. (2016). Enactments concerning the Irish language, 1922-2016. *Dublin University Law Journal 39*(2), 449-466.

[4] Sutskever, I., Vinyals, O. & Le, Q.V. (2014). Sequence to sequence learning with neural networks. [Paper presentation]. *Advances in Neural Information Processing Systems*.

[5] Bahdanau, D., Cho, K. & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. [Paper presentation]. *The 3rd International Conference on Learning Representations*.

[6] Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. [Paper presentation]. *The 2016 Conference on Empirical Methods in Natural Language Processing*.

[7] Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivavasta, A., & Judge, J. (2015). Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish. *The 4th LRL Workshop: "Language Technologies in Support of Less-Resourced Languages*.

[8] Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Peréz-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G. & Tyers, F.M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation 25*(2), 127-144.

[9] Dowling, M., Lynn, T., Graham, Y. & Judge, J. (2016). English to Irish Machine Translation with Automatic Post-Editing. In *Proceedings of the 2nd Celtic Language Technology Workshop* (pp.42-54).

[10] Dowling, M., Lynn, T., Poncelas, A. & Way, A. (2018). SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages* (pp. 12-20).

[11] Arcan, M., Lane, C., Ó Droighneáin, E. & Buitelaar, P. (2016). IRIS: English-Irish machine translation system. [Paper presentation]. *The 10th International Conference on Language Resources and Evaluation.*

[12] Torregrosa, D., Pasricha, N., Masoud, M., Chakravarthi, B.R., Alonso, J., Casas, N. & Arcan, M. (2019). Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII, Volume 2: Translator, Project and User Tracks* (pp. 125-133).

[13] Shaw, S. (2013). *Shallow transfer rule-based machine translation and the Irish-English language pair*. [Final year undergraduate project report]. Trinity College, Dublin.

[14] Dowling, M. (2015). *Rule Based Machine Translation for the Irish English pair using Apertium.* [Final year undergraduate project report]. Trinity College, Dublin.

[15] O'Regan, J. (2017). *English-Irish Machine Translation*. [M.Phil. thesis]. Trinity College, Dublin.

[16] Hamman, E. (2019). *Evaluating Google Translate for the English Irish language pair and Implementing Online Translation Checking.* [Final year undergraduate report]. Trinity College, Dublin.

[17] Uí Dhonnchadha, E. (2009). *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. [Ph.D. thesis]. Dublin City University.

[18] Eriguchi, A., Tsuruoka, Y. & Cho, K. (2017). Learning to Parse and Translate Improves Neural Machine Translation. [Paper presentation]. *The 55th Annual Meeting of the Association for Computational Linguistics.*

[19] Bernardinello, G. (2019). Morphological Neural Pre- and Post-Processing for Slavic Languages. In *Proceedings of Machine Translation Summit XVII, Volume 2: Translator, Project and User Tracks.*

[20] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318).

[21] Denkowski, M. & Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

[22] De Souso, S.C.M., Aziz, W. & Specia., L. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 97-103).

[23] Fuji, M., Hatanaki, N., Ito, E. Kamei, S., Kumai, H. Sukehiro, T., Yoshimi, T. & Isahara, H. (2001). Evaluation method for determining groups of users who find MT "useful". In *Proceedings of Machine Translation Summit VIII.*

[24] Moorkens, J., Castilho, S., Gaspari, F. & Doherty, S. (2018). *Translation Quality Assessment: From Principles to Practice.* Springer.

[25] Snover, M., Dorr, B., Schwartz, L. Micciulla, L. & Makhoul, J. (2006). A study of translation edit rate with target human annotation. In *Proceedings of the Association for Machine Translation in the Americas* (pp. 223-231).

[26] Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Prytz, L.-I. & Wester, M. (2012). Eye Tracking as a Tool for Machine Translation Error Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation.*

Chapter 12
Information Retrieval

12.1 What is Information Retrieval?

Information Retrieval (IR) is a method for accessing information from a large collection of data sources. This data may take many forms such as text documents, images and audio-visual files. The storage format for this data will vary from archived directories, local repositories, online repositories and internal databases to web domain repositories and so on.

As an area of research, IR tends to focus on improving the accuracy with which keyword searches return relevant and useful results. Results are usually ranked according their perceived relevance to the search. There are various types of IR applications, and we benefit from a few specific types in our daily lives. For example, on a web search engine, the keyword search 'evening classes in Dublin' could return a list of pages providing information on adult classes, community evening classes, university-hosted evening courses, etc. A business database search for '2015 supplier reports' would expect to retrieve all documents related to reports relating to suppliers for the year 2015.

The enormous increase in the amount of online text available and the demand for access to different types of information have, however, led to broader range of IR-related areas that go beyond simple document retrieval. Some examples include question-answering, topic detection and tracking, summarisation, multimedia retrieval e.g. image, video and music), software engineering, chemical and biological informatics, text structuring, text mining and genomics [1].

12.2 Why is it important?

As the amount of data available online and in private archives is increasing, the need for accurate search technology is becoming more apparent to allow users to access the information they need. The need for reliable IR systems is not restricted to a general world-wide web search, however. Individual websites, with a large number of web pages or large amounts of archived data require accurate search engines.

In addition, as any business or organisation's database of documents grows, the need for tools to easily browse this data also grows. The need for reliable IR can vary based on the type of content an organisation holds or archives, and the purpose of the retrieval efforts. For example, a business may need to accurately search through customer records in order to generate or analyse reports. A library will have records of books, publications, journals, newspapers, etc. A broadcaster will have an archive of interviews, documentaries, reports, subtitle files and so on.

In the context of the Irish language, there are many situations in which improved IR systems are needed. Tuned IR systems can contribute to more efficient data management for business organisations, government online and internal research, academic online and archival research, media institutions' archive retrieval of audio-visual content, along with broader online research of multimedia files.

12.3 How does it work?

A basic Information Retrieval (IR)/Search Engine is relatively language-independent, focusing only on indexing and ranking of pages and documents according to the relative frequency of the terms used in the keyword search in the collection of documents being searched. However, reliable search tools are language-dependent. That is, their design considers linguistic information of the language in the documents or web pages being searched. The success of current IR research is as a result of the integration of natural language processing methods that allow the system to be more language-aware. These methods help to enhance representation and understanding the queries and the documents to ensure improved matching.

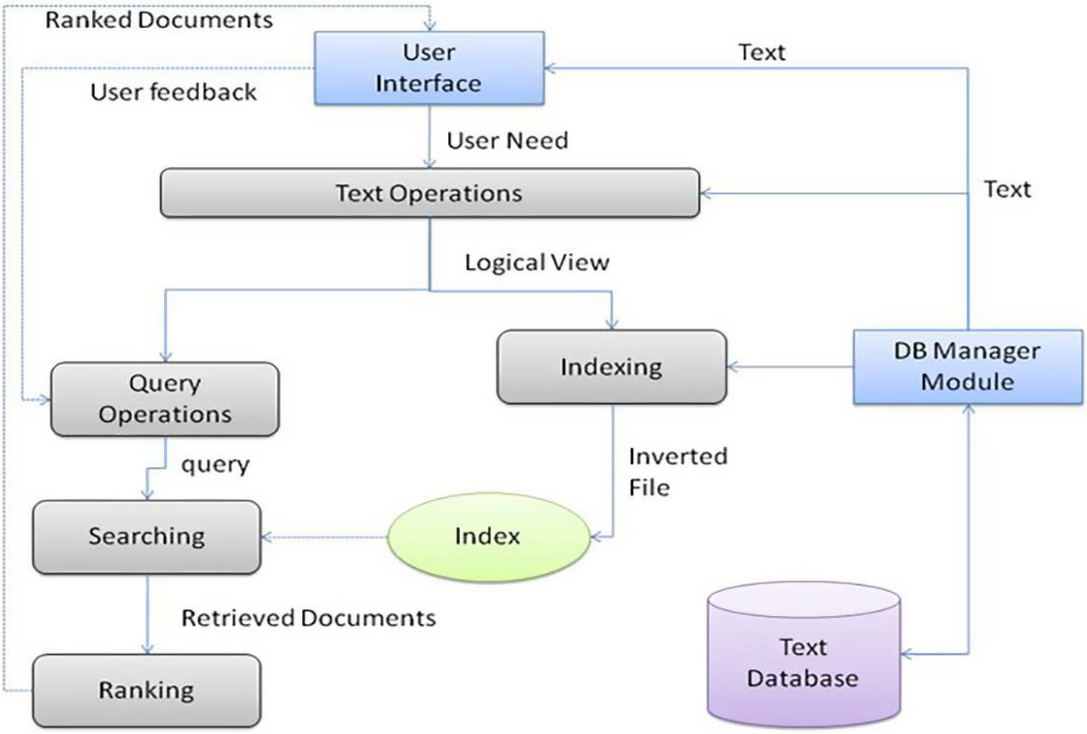


Figure 12.1: Example Overview of an Information Retrieval Process [2]

Highly efficient search engines exist for well-resourced languages such as English, German and French, for example. These search engines have seen significant improvements in recent years in terms of the percentage of relevant web pages or documents retrieved through keyword searches. These advancements in IR technology are restricted, however, to well-resource languages. IR technology tailored to low-resourced languages such as Irish are usually rare or non-existent. While web search engines such as Google may provide an Irish language interface, their search tools do not demonstrate an underlying deeper understanding of the language to bring the quality of results in line with that of other better supported languages.

12.4 What has been done for Irish?

There have been some limited localised efforts to building IR systems for small websites (e.g. www.gaeilge.ie by Foras na Gaeilge). Yet, to date, no extensive research has been carried out or reported on regarding the application or evaluation of current state-of-the-art IR methods to the Irish language; or indeed the development of new IR methods tailored to the Irish language. This chapter therefore highlights all areas of information retrieval R&D that requires attention in the near future.

What is important to keep in mind, is that currently, Irish speakers are using online and offline search tools that have been tailored and optimised for searching through English language digital content. These tools are thus limited in their ability to correctly process Irish text or audio files, and subsequently produce broadly inaccurate and incomplete results.

12.5 Recommendations

Research and development of Irish IR systems will mark a significant advancement in managing and accessing Irish language resources that will significantly benefit members of the public, TV and Radio corporations, businesses and governments alike. The following are key areas recommended as priority targets over the next 5-10 years.

- **Text search and retrieval tuned to the Irish language**

High-quality IR systems require tuning to the specific language of a query. Many of the underlying language technology resources required for basic Irish IR tuning are discussed in earlier chapters. Here we highlight some of those areas that will need to be adapted to IR-specific research. It is worth noting, however, that successful IR engines rely on much more than the integration of the tools listed below:

- Tools such as stemmers or lemmatisers are used to find the stem of a keyword and assist an IR engine in retrieving documents with other words morphologically related to that stem, i.e. instead of only restricting the keyword search to the inflected form. For example, a search on the keyword *teicneolaíocht* should also return results containing pages that may only have the inflected forms *teicneolaíochta* or *teicneolaíochtaí* present (See Chapter 5 for further reference).
- Similarly, semantic information (meanings of words) can be drawn from lexical semantic networks such as WordNet to inform a search engine of relationships between words, such as synonymy [3]. For example, a top result for the English keyword search 'Irish language technology' is a web page describing a new language learning 'app'. While the word 'technology' does not feature in the resulting, highly ranked document, the English IR system is intelligent enough to recognise the semantic relationship between 'app' and 'technology'. The LSG Irish Language Semantic Network (see Chapter 4) would be an ideal contributor of this kind of data for Irish IR systems.
- In addition, parsers are sometimes used to assist IR methods, by providing an additional layer of information regarding the structure of the text in documents being searched, or for extracting the important elements of a question-based search [4]. The parsers discussed in Chapter 5 would prove useful in identifying and analysing the structure of text for the improvement of Irish-language optimised search engines. This approach would also apply to speech-based queries.

Any research carried out first on text retrieval can then be extended to searching through multimedia archives that have been manually created or have automatically extracted transcripts attached to them. This type of technology would be invaluable for organisations such as Irish-language TV stations, radio stations, digital humanities archiving and research, and so on. In some cases, systems may not have to be built from scratch but instead existing open-source frameworks such as Terrier could be adapted if Irish-specific NLP tools were incorporated appropriately.

- **Information retrieval results summarisation**

When a list of results (e.g. documents) are presented to a user, the user needs assistance in deciding which documents are most relevant to their search, without examining each document individually. A solution often provided by search engines is the use of snippets, which usually contain a document title and a 2 to 3 line summary of the document. In general, the techniques focus on identifying a select few sentences of the document that are deemed representative of the content. This summarisation task can sometimes prove difficult in IR development as it relies on smart Natural Language Processing (see Chapter 5) and Natural Language Generation (see Chapter 6) techniques specific to the language in question, such as language modelling, generation or parsing.

Summarisation of Irish documents that are found in search results would be an invaluable contribution towards improving the search facilities of not only online content, but also internal archived legacy content that can often appear too large and unwieldy to process easily.

- **Cross-lingual IR**

Cross-lingual IR is a branch of research that combines machine translation (MT) and information retrieval techniques. Much of the current focus in this area is the attempt to overcome the lack of digital data available for lesser-resourced languages, by providing additional relevant results of a keyword search in other better-resourced languages.

For example, the Basque IR system 'Zientzia' [5-8] allows a user to search an archive of Basque-language scientific articles by entering Basque keywords. The first set of top-ranked Basque-language documents are presented to the user. In addition, results in both English and Spanish are also returned, thanks to the integration of Basque-English and Basque-Spanish MT systems. This approach widens the accessibility of data to a user who chooses to search in Basque only. The system also indexes other science websites and provides them as secondary result options. Similarly, the online search engine 'elebila™' is a cross-lingual web-based IR system that allows a user to search in a choice of four languages (Basque, English, Spanish, French) and retrieve all relevant results. There is also an option to extend the search to include synonyms and inflected synonym forms. Elebila provides users with plug-ins for Word or Open Office as well as a web page plug-in.

Given that the amount of data available online in Irish is still relatively limited, it is strongly recommended that research is conducted into the development of a cross-lingual IR system such as those available for

Basque speakers. Such a system that would integrate current and future Irish-English machine translation technologies (see Chapter 11) and would allow a user to search for keywords in Irish and to retrieve not only the highly ranked results for online Irish documents, but also English documents containing relevant information that may also prove useful to the user.

- **Irish spoken content retrieval**

Once a reliable and robust automatic speech recognition (ASR) system (see Chapter 9) is made available for Irish, the resulting ASR transcripts can be indexed to be used within an IR system, thus enabling searches through audio and video archives.

Initially, the transcripts can be considered as a variation of textual input for the IR system, and its initial customisation for the Irish language can be immediately implemented (e.g. through the use of stemmers and stop word lists [9]. Further cross-lingual IR can be tested to expand the search to the multimedia archives in other languages. It is important to notice that even ASR transcripts with 100% recognition quality still differ from traditional text, as people use shorter sentences, interruptions, disfluencies, and so on, while speaking.

Therefore, further Spoken Content Retrieval system development and quality improvement will require investigation into differences between spoken and written language, and potentially require specific tools-tuning. For example, sentence parsing for spoken content might be more similar to parsing unedited social media content than to parsing regular newspaper text, due to the presence of disfluencies and incomplete phrases.

References

[1] Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Damais, S., Fuhr, N., Harman, D., Harper, D.J. & Hiemstra, D. (2003). Challenges in information retrieval and language modelling. *ACM SIGIR Forum 37*(1), 31-47.

[2] Buscaldi, D. & Rosso, P. (2011). Explicit query diversification for geographical information retrieval. In *Proceedings of the 33rd European Conference on Information Retrieval* (pp. 73-80).

[3] Rosso, P., Ferretti, E., Jiménez, D. & Vidal, V. (2004). Text Categorization and Information Retrieval Using WordNet Senses. In *Proceedings of the Second Global Wordnet Conference* (pp. 299-304).

[4] Maxwell, T.K., Oberlander, J. & Croft, B.W. (2013). Feature-Based Selection of Dependency Paths in Ad Hoc Information Retrieval. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 507-516).

[5] Saralegi, X. & López de Lacalle, M. (2010). Estimating Translation Probabilities from the Web for Structured Queries on CLIR. In *Proceedings of European Conference on Information Retrieval* (pp. 586-589).

[6] Saralegi, X. & López de Lacalle, M. (2010). Dictionary and monolingual corpus-based query translation for Basque-English CLIR. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (pp. 1353-1358).

[7] Saralegi, X. & López de Lacalle, M. (2009). Comparing different approaches to treat translation ambiguity in CLIR: structured queries vs. target co-occurrence based selection. In *Proceedings of International Workshop on Database and Expert Systems Applications* (pp. 398-404).

[8] Saralegi, X. & Alegria, I. (2007). Similitud entre documentos multilingües de carácter científico-técnico en un entorno Web. *Procesamiento del Lenguaje Natural 39*, 71-78.

[9] Uí Dhonnchadha, E. & van Genabith, J. Scaling an Irish FST morphology engine for use on unrestricted text. (2006). In A. Yli-Jyrä, L. Karttunen & J. Karhumäki (Eds), *Finite-State Methods in Natural Language Processing 5th International Workshop, Revised Papers*. Berlin: Springer International Publishing, pp. 247-258.

Part III

Applications

Chapter 13
Applications for Public Use

13.1 What is it?

We are living in a transformative period in human history. Speech and language technologies are now woven into the fabric of everyday life so smoothly that we are often unaware of their extent. As emphasised in the introduction (Chapter 1) they are part of how we live, work, play, communicate, seek information, learn and interact with services and our environment.

A goal of the Digital Plan is to develop applications that will enable the Irish language to be integrated into a daily and increasingly digital life. It is essential that Irish can be used in all domains, e.g. in commercial contexts, legal contexts, leisure etc. When the domains of usage shrink, the language loses prestige: a technology gap means that people have to default to English in more and more of their daily activities, accelerating the rate of language attrition. The corollary is that provision of sophisticated speech and language technology brings untold new opportunities to establish Irish firmly in all spheres of life.

Digital applications for the public will include a wide range of systems, whose complexity will grow as technology advances. These include:

- *standalone applications* that exploit linguistic modules, e.g. a grammar checker in your word processor
- *complex systems* that directly employ the core technologies, e.g. a dictation system exploiting speech recognition
- *combined complex systems*, e.g. speech-to-speech translation systems, which integrate speech recognition, machine translation and speech synthesis in a second language. This can also include multi-language captioning so that the user can both listen to and read a speech in Irish as it is being delivered in another language
- *interactive*, *virtual-reality-based systems*, where speech and language technologies live inside an embodied agent, e.g. a robot, with which the user interacts for particular activities. Communication is further enhanced by gesture and facial expression

There is vast potential for ways in which speech and language technologies can be embedded in the public domain but this chapter will simply give a flavour of such potential. As two vital areas, education and access, are explored in Chapters 14 and 15, discussion here focuses on other domains.

13.2 Why is it important and for whom?

The development of these applications is important for the whole language community, including those who work and live in an Irish-language setting, learners and those who, for one reason or another, wish to engage with the language. Here, some more futuristic applications are discussed, along with more established technologies, that will affect different aspects of our lives.

13.2.1 Spheres of daily activity

Speed, efficiency and convenience are a main driver for the take-up of many technologies and creative *integration* of the speech and language components is key to most applications.

At Home: smart environments. We're living in homes that are becoming 'smarter' by the day. Increasingly we control appliances and our home environment using technology, often voice-controlled: e.g. searching for TV/radio programmes; controlling lights, security, heating, entertainment systems etc. and these can often be done remotely using a smartphone app. Automatic subtitling of radio or other audio content helps make it accessible, and is particularly helpful for those who might have difficulty hearing. Multitasking means that we may be communicating verbally while our hands or eyes are busy. For example, people already speak to their GPS while driving the car, or voice assistants while cooking dinner.

At Work: virtual assistants. Standalone NLP-based modules are already widely deployed for content composition, including spelling and grammar checkers, which are helpful for proofing and editing. These will need refinement and extension (see more in §13.4).

As new core technologies come on stream, content composition will be further enhanced, e.g. dictating text, prooflistening to complement visual proofing, and eventually accurate text summarisation tools will speed up the processing of large amounts of information.

With combined complex systems the range of useful applications keeps growing. For example, by including machine translation technology, searches conducted through Irish can retrieve (in Irish) information available in other languages. With speech technology this information could be sought and delivered by voice.

Social and Leisure Activities: social networking using mobile devices increasingly links us to our friends, initiates social connections, informs us of what events are happening, what friends are doing and suggests who we might like to get in touch with. Intelligent agents are increasingly being used to organise our social calendar, e.g. booking tickets to events or making reservations.

Other leisure activities, such as visiting a museum/art gallery will increasingly entail interacting with exhibits, e.g. embodied agents representing historical characters from the past will talk and answer your questions (see Chapter 10). Current versions are relatively simple and with advances in Augmented Reality, stories from our everyday environment can come alive. Imagine walking around the Blasket Islands and experiencing life on the island when there was a thriving community!

Interactive multimodal games have taken over as a major form of leisure for all ages and offer many opportunities to have fun through the medium of Irish. Currently, popular games include using voice assistants, e.g. to host family game-show trivia games or children's party games, like musical chairs. Children, as 'digital natives' who have not known a world without technology, will expect 'intelligent' toys that listen and talk back to them. As they get older, coding activities, e.g. CoderDojo[1] and Hour of Code[2] through Irish should encourage the younger generation to creatively fuse their language with fun game development.

Interactions with the State and public bodies: speech and language technologies empower the State and pubic bodies to meet their obligations under the Official Languages Act to offer all citizens opportunities to interact through Irish. Machine translation is a key technology in that it would ensure bilingual versions of all documents. Speech and language technologies are both central in allowing easy access to information, forms, etc. and can allow tasks to be carried out either by voice or by text.

The State is also obligated to ensure that information can be readily accessed by all, including those with disabilities (see Chapter 15). Crucial technologies here include text-to-speech, for those with visual and vocal difficulties; speech-to-text to subtitle audio streams for those who cannot hear; and text summarisation tools, which are particularly helpful for those with attention difficulties, to e.g. navigate complex documentation.

Health: the face of health provision is rapidly changing. Increasingly, voice assistants are being used to respond to requests for information about health. There is a growing field of use of robots as assistants, e.g. for the elderly or disabled, to help maintain independence, to provide companionship, to set reminders regarding medication, appointments etc. For those with motor or other disabilities, such robots can carry out specific activities as a response to voice commands, e.g. as simple as boiling a kettle and making a cup of tea, that would facilitate independent living.

There is also increasing use of systems (e.g. phone apps) where the individual's speech and language is monitored for changes in physical or mental health. For example, the monitoring of the voice prosody (see Chapter 2.4.3) is being used to detect signs of depression, serving as an early warning system to ensure that therapists' interventions are provided in a timely fashion.

13.2.2 Building and connecting communities: technology for good

Building communities: new communities arise spontaneously online through fora, social media platforms etc. that bring together people with shared interests, e.g. music or sport. The platforms are predominantly in English but are now being translated into Irish (see 13.4.3 below). The increasing power of text and speech tools will facilitate the organic formation of interest groups who simply use the language, not necessarily as the focus of interaction, but as their medium of communication. Whereas in the past networks were inherently local, the younger, tech-savvy generation are now empowered to find like-minded individuals anywhere in the world and build new networks in a way that is not initiated or controlled from the outside.

[1] https://coderdojo.com
[2] https://hourofcode.com/ie

Community connecting: as discussed in Chapter 1, the Irish language community is dispersed in geographically remote Gaeltachtaí and in small pockets outside the Gaeltacht, in Ireland and abroad. TG4 and Raidió na Gaeltachta have already transformed the level of linkage between the Gaeltacht populations, and connected them in turn to the wider language community. Speech and language technologies are the next frontier, promising a diversity of enhanced platforms, e.g. linked to embodied agents or Augmented Reality, that have enormous potential to enlarge and strengthen these networks, empowering the more isolated individuals and groups through a wider community identity as 'pobal na Gaeilge', sharing in a rich cultural heritage.

13.3 How does it work?

As discussed above, a wide variety of applications that work in diverse ways are needed. The Digital Plan offers, not just the possibility of catching up with technologies that already exist for English but the opportunity to explore creative new applications that specifically answer to the context of Irish in the 21st century.

13.4 What has been done to date?

A number of applications for public use are already available and more are under development. Examples below include **text-based tools**, e.g. for document proofing and readability, predictive text for mobile and a dictionary look up. **Speech-based facilities** include a web page, a webreader and an Android app, which enable more widespread access to audio read out of any text online. A further important area is **localisation**, which involves customisation of existing platforms and applications for the Irish language and context.

13.4.1 Text-based Tools

Document Proofing Tools. These include spellcheckers, grammar checkers, style checkers and are currently the most widely used language processing tools. For the major languages these are pre-installed in word processors e.g. MS Word, OpenOffice etc. and in email applications e.g. MS Outlook.

Gaelspell Spellchecker[3] is a free, open source spellchecker, developed at Saint Louis University, Missouri[4]. It can be used on any platform (Linux, Mac, or Windows) and there are installable packages for open source applications, like Mozilla Firefox, Mozilla Thunderbird, OpenOffice, and LibreOffice. It has also been packaged for other applications, including Microsoft Word and Mac OS[5] X - a good example of how opensource packaging can yield impact beyond the original development.

Microsoft Word Irish spellchecker, developed at Trinity College Dublin, is available for Microsoft Office for use with Microsoft Word. It can be downloaded from Microsoft Irish Proofing Tools[6]. However, this spellchecker has not been updated since its initial release in 2002, and its dictionary now requires updating[7].

An Gramadóir is a free, open source, rule-based Irish grammar checker, also developed at Saint Louis Univerisity, Missouri. It uses part-of-speech tagging and 2,800 pattern-matching rules to highlight the most common grammatical errors in Irish. It has been integrated into end-user-facing applications, such as Microsoft Word (via the commercial packages 'Ceart' and 'Anois' by Cruinneog) and LibreOffice (via a collaboration with Dúrud and COGG)[8].

A limitation is that it does not perform deep (syntactic) linguistic analysis, and consequently some grammatical errors cannot be detected, and there are occasional 'false positives'. It could be improved through the integration of more advanced NLP tools, e.g. a rule-based or statistical chunker, or a dependency parser (see Chapter 3).

Readability tools draw the writer's attention to how accessible and easy to understand their document is. It adds to spelling and grammar checkers in that it offers suggestions on how the text might be clarified or simplified to make it easier to read. Although long in use for English [1], e.g. the well known *Grammarly* [2], these tools are not yet available for Irish. Early research towards readability tools has been carried out in Trinity College Dublin [3]. However, as with most speech and language technologies, the linguistic underpinning is predominantly based on English and the resources discussed in Part I are prerequisites to provide for the needs of Irish.

Predictive text for mobile: *Adaptxt* is an application for mobile phones, designed to improve text entry by making it faster and error-free. It predicts the word being typed as well as the following word, and adapts continuously to the user's vocabulary and writing style.

[3] https://github.com/kscanne/gaelspell/commits/master
[4] https://cs.slu.edu/~scannell/
[5] https://cruinneog.com/
[6] https://www.microsoft.com/ga-IE/download/details.aspx?id=52668
[7] http://www.scriobh.ie/Page.aspx?id=18&1=2
[8] http://www.scriobh.ie/Page.aspx?id=78&l=1

At nascanna.com[9], users can set Adaptxt to include the Irish dictionary and make it the primary keyboard in order to avail of Irish predictive text and spellchecking.

Dictionary plug-in: a dictionary plug-in (or add-on) allows someone reading a web page to directly access dictionary information without leaving that web page. Plug-ins are being made available from Irish dictionary websites: Figure 13.1 illustrates the teanglann.ie plug-in in use with the online newspaper tuairisc.ie. This valuable tool may be enhanced by further features, e.g. to handle inflected forms (see Chapeter 2.5.1).

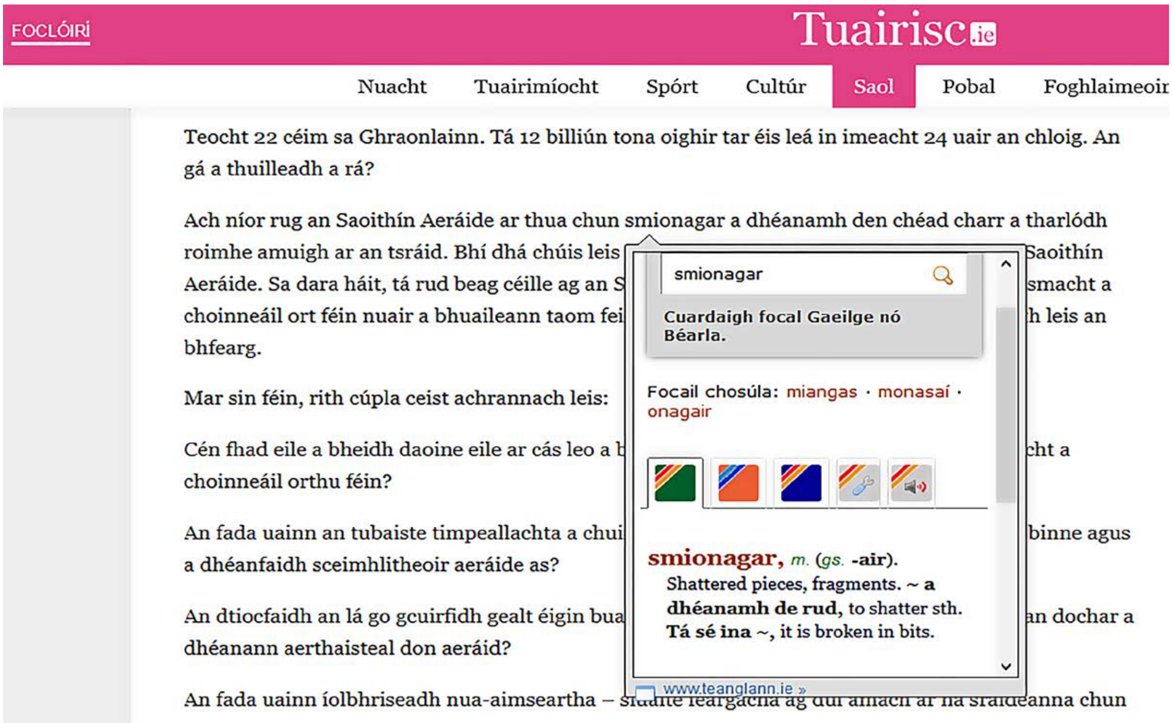
13.4.2 Speech-based facilities

Speech output of any online text is now provided through different routes. On the ABAIR web page, www.abair.ie, a beta version of a speech recognition system, ABAIR_ÉIST, was recently launched on the same site, allowing text composition from speech.

The ABAIR website provides access to the range of synthetic voices now available for the Irish dialects as well as serving as a portal to the ongoing research on speech technology for Irish. This website is widely used in Ireland and abroad (see Chapter 8). This portal also provides access to further applications geared at accessibility and educational use.

An ABAIR Android app has also been designed, providing easy access to speech output on smartphones/tablets . The current implementation is for the Connemara voice and the other dialects will soon be included.

A further facility is an ABAIR plug-in, which has been developed to make it easier for the public to access the spoken output on any device. The individual user can also download it to their browser, and activate it when reading any Irish text online. A webreader is a separate facility that can be built into Irish language web pages by their developers[10]. It acts as a customised reader where the choice of dialect, the speed of output, etc. is under user control and individual words may be highlighted as they are read aloud (Chapter 8). In the case of dictionary entries where the headword is replaced by a ~ in the examples, the missing root is filled in when the word is spoken aloud. For example, the future form of the entry 'tóg' (displayed as ~faidh) is pronounced as *tógfaidh*.



Teocht 22 céim sa Ghraonlainn. Tá 12 billiún tona oighir tar éis leá in imeacht 24 uair an chloig. An gá a thuilleadh a rá?

Ach níor rug an Saoithín Aeráide ar thua chun smionagar a dhéanamh den chéad charr a tharlódh roimhe amuigh ar an tsráid. Bhí dhá chúis leis … Saoithín Aeráide. Sa dara háit, tá rud beag céille ag an S… smacht a choinneáil ort féin nuair a bhuaileann taom fei… h leis an bhfearg.

Mar sin féin, rith cúpla ceist achrannach leis:

Cén fhad eile a bheidh daoine eile ar cás leo a b… cht a choinneáil orthu féin?

An fada uainn an tubaiste timpeallachta a chui… binne agus a dhéanfaidh sceimhlitheoir aeráide as?

An dtiocfaidh an lá go gcuirfidh gealt éigin bua… an dochar a dhéanann aerthaisteal don aeráid?

An fada uainn íolbhriseadh nua-aimseartha – sl… ag dul amach ar na sráideanna chun

smionagar

Cuardaigh focal Gaeilge nó Béarla.

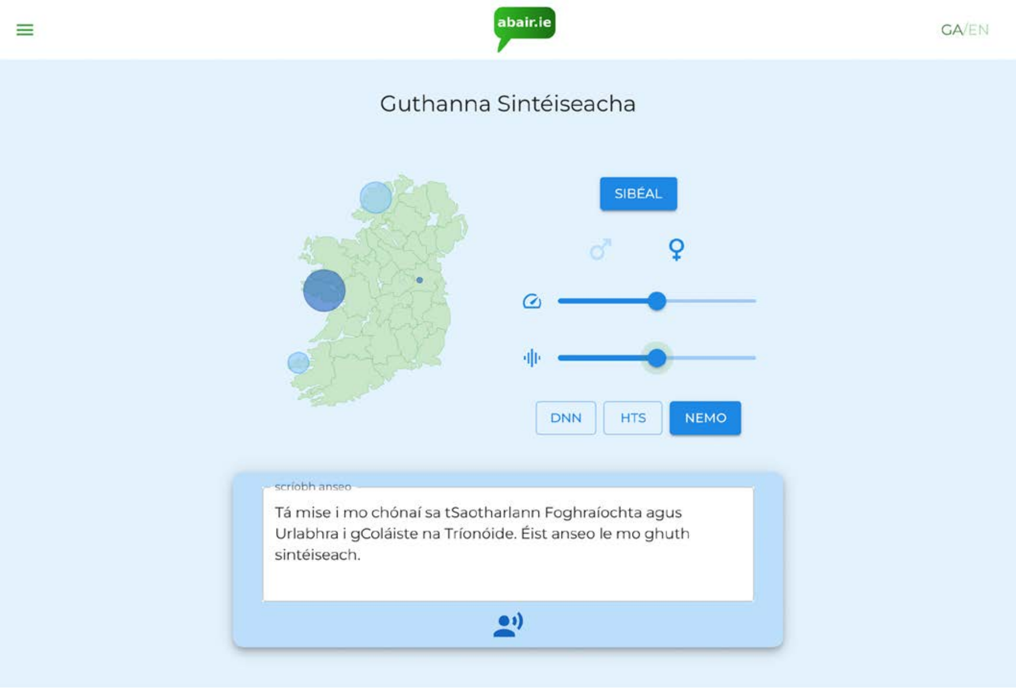
Focail chosúla: miangas · monasaí · onagair

smionagar, *m.* (*gs.* **-air**). Shattered pieces, fragments. ~ **a dhéanamh de rud,** to shatter sth. **Tá sé ina ~,** it is broken in bits.

www.teanglann.ie »

Figure 13.1: teanglann.ie plug-in on the tuairisc.ie website

[9] http://www.nascanna.com/NascannaCom/Feidhmchlar.aspx?id=90
[10] https://www.abair.tcd.ie/en/accessibility/



Tá mise i mo chónaí sa tSaotharlann Foghraíochta agus Urlabhra i gColáiste na Tríonóide. Éist anseo le mo ghuth sintéiseach.

Figure 13.2: www.abair.ie web page

The webreader is particularly important for public bodies that are obligated by law to provide information through Irish[11] and furthermore to ensure accessibility to all, including those with disabilities (cf. *Disability Act 2005*, §26 (1)(a)). A number of synthesis-based applications primarily geared towards education and access are described in Chapters 8, 14 and 15. Note, however, that many of these are of value to the broader Irish-language public. For example, robot receptionist chatbots, based on spoken dialogue systems (Chapter 10), will feature more and more in public spaces, e.g. science galleries, museums, etc.

13.4.3 Localisation of Everyday Applications

Localisation is the process of adapting a product or content to a specific region or market. There are two aspects involved: firstly, translation to the local language and secondly the modification of content to consider cultural and political sensitivities, date, time and currency formats, colours and sounds, social factors and legal requirements of the country for which they are being adapted.

Part of the objective of the Digital Plan is to have Irish used in all domains. So much of the digital world is developed for English and a different cultural context. Having Irish language versions of the digital platforms we use daily extends the activities and areas in which it is natural to use the language, contributing to its prestige and widespread use.

An Ríomhacadamh is an organisation of volunteer developers, translators, and technologists whose mission is to provide the Irish-speaking community with the software they need to operate online fully through Irish. They have to date localised GMail, Twitter, Mozilla Firefox, OpenOffice, LibreOffice, Code.org, Scratch, WhatsApp, Skype, and dozens of other applications, websites, and games. They also maintain an important Style Guide for Irish software localisation that members are expected to follow to ensure consistency in terminology, grammar, register, etc. across all products available in Irish[12] .

Other multinational corporation such as Microsoft and Airbnb have also commissioned professional translation services to localise their applications to Irish. Note, however, that support for a localised version of MS Office has halted since MS Office 2010. For open-source products, ongoing support for the Irish locale is provided by OpenOffice.

[11] https://www.chg.gov.ie/app/uploads/2020/07/acht-na-dteangacha-oifigiula-2003.pdf
[12] https://riomhacadamh.wordpress.com/stil-ti/

A good example of localisation which includes adaptation to the local context is the video game *Minecraft*. A crowdsourcing approach was taken to translating the gaming content[13] and linked geographical data (see Chapter 4) from Ordinance Survey Ireland (OSI) was integrated into the game enabling Minecraft players to explore a virtual Ireland[14]. Other internationally-renowned games development companies, e.g. Romero Games, are also beginning to localise their games for an Irish-speaking audience[15].

As pointed out in Chapter 1, technologies for Irish are not simply a matter of taking a technology developed for English and adjusting it superficially for Irish. Just as when we build new purpose-built technologies for Irish (Parts I and II), we need to take account of the linguistic structure of Irish and the context of its users when localising existing products and applications. Thus, for example, localising software for certain mobile phone applications would require adding a predictive text tool that recognises words and obeys the linguistic structures of the local language.

All the resources, tools and technologies emerging from the Digital Plan will contribute to various aspects of localisation. Machine translation is a particularly important technology that will facilitate widescale localisation into the future.

The EU's eTranslation system is available for use for free by those in the Irish public sector and Irish SMEs (small-to-medium enterprises).

13.5 Recommendations for future work

Creative minds will lead to many applications that are not foreseen at this point, and every incremental step of the Digital Plan will inspire innovations. Building applications at the same time as core technologies are being developed will ensure that they are fit for purpose and relevant to the context in which they will be used.

To ensure that the outputs of the Digital Plan reach the public who should benefit from it, the following are recommended:

- The priorities for building applications should be based on public need.
- The Irish language community as the central partner needs to be involved in the process of developing applications from the outset. Their input may take several forms, such as determining priorities or providing support and testing for applications as they are developed (this is further discussed in the Conclusions).
- The development groups need support so that their outputs are well-packaged, easily installed, maintained and updated to be compatible with the changes to operating systems (e.g. updates to Windows), programming languages (e.g. python updates), etc. A facilty for ongoing user support is also recommended.
- Support for interfacing tools and applications to be compatible as third party tools with widely used platforms, such as Microsoft 365. This also will require maintenance and technical support. This would be an ideal opportunity for companies to showcase their commitment to the local community as part of their Corporate Social Responsibility.
- Basic language tools and applications for the public must be embedded within and shipped as a basic component of digital packages provided by the computing industry in Ireland.
- Government support to ensure that these tools and applications are a basic provision for all in the public service, including the civil service, schools, and third level institutions.
- Government support to ensure widespread awareness of the technologies that are available and to encourage public take-up. The Government can also be proactive in interactions with the large technology companies to ensure that the position of the Irish language as the first national language is appreciated and promoted.

References

[1] Kincaid, J.P., Fishburne, R.P., Rogers, R.L. & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel. Institute for Simulation and Training, University of Central Florida.

[2] Grammarly, Inc. (2014). Grammarly. www.github.com/grammarly

[3] Ó Meáchair, M.J. (2020). The Creation and Complexity Analysis of a Corpus of Educational Materials in Irish (EdaGA). [Ph.D. thesis]. Trinity College, Dublin.

[13] https://crowdin.com/project/minecraft/ga-IE#
[14] https://www.osi.ie/creating-ireland-in-minecraft
[15] https://tuairisc.ie/daithi-an-dainseir-ar-fail-an-athuair-don-ios-agus-gaeilge-bhrea-aige/

Chapter 14
Educational Applications and Computer-Assisted Language Learning (CALL)

14.1 Introduction

One of the greatest dividends and the most concrete application and manifestation of the digital resourcing of Irish is likely to be in the sphere of language education. The potential exists to make the Irish language an integral part of the burgeoning technological era, participating in the vision articulated by the Department of Education and Skills, to 'realise the potential of digital technologies to enhance teaching, learning and assessment' [1]. The outputs of the Digital Plan will contribute in numerous ways. Empirical analyses of quantitative data using state-of-the-art tools promise a fresh take on the structures of Irish and knowledge of the processes whereby native speakers and second language learners acquire it. This will provide scientific underpinnings for curriculum design and materials development for teaching and assessment. Furthermore, the core technologies of the Digital Plan will pave the way for exciting and innovative Computer-Assisted Language Learning (CALL) applications, which have the potential to revolutionise Irish language teaching/learning. They will also facilitate Irish language usage in a wide range of other educational activities.

CALL can range from the simple to the very complex. At one end it can entail relatively simple programmes that allow training on one specific aspect of the language, e.g a mini-game with attractive visuals which can be an effective aid for the learner. At the other end of the scale, intelligent CALL applications entail much more complex learning platforms, which draw on many technological innovations. The latter include multimodal, speech-enabled, interactive platforms incorporating natural language processing (NLP) and speech components - where the learner is an active participant in a specially-designed, structured game, or, where they enters a virtual reality immersive environment where all the interactions are in Irish. These platforms exploit sophisticated gaming technology, embed knowledge of the language and marry these to serious language learning goals. These are identified for their potential for major impact in the Irish-language learning context and particular focus is placed on them here.

Developing these complex platforms will entail multidisciplinary teams, including user interface/gaming designers and software developers (§14.5 below). However, the crucial element is the pedagogical and linguistic input provided by CALL specialists, educationalists and practitioners to ensure that these applications are based on (i) sound pedagogical principles, (ii) an understanding of the language structures and the acquisition process and (iii) clarity regarding the intended end users and their specific needs. In this endeavour, vital input will be derived from the linguistic and other knowledge bases of Part I, including the analysis of learner corpora (see Chapter 2).

14.2 Why is CALL important in the Irish context?

Since the foundation of the State, it has been appreciated that education is key to the maintenance of Irish as a spoken language in Ireland. Although Irish is a compulsory school subject until school-leaving age and thus has a large learner community, and although the population at large is positively disposed to the language [2], this often does not translate into enthusiastic learners and successful language learning outcomes [3]. The trend in recent years has been towards a decline in standards [4], and it is regrettable that after 13 years of Irish language instruction, school leavers are often unable to carry out a simple conversation in Irish.

The single biggest problem is that most learners do not have enough exposure to the language and have very little direct contact with native speakers, given that Irish is spoken as a community language only in Gaeltacht areas. The dearth of native speaker models of the language in many learners' environment and the lack of social contexts where true interactions can take place in the language (the optimal conditions for language learning) are obstacles to effective language learning. Teachers, typically being second language learners themselves, can sometimes feel insecure about their own level of competence, and they often feel isolated and unsupported with the responsibility of language maintenance that rests on their shoulders [5]. Parents can also feel ill-equipped to provide support with homework, a problem which is particularly acute for immigrants.

Many commentators agree that the acquisition of native sounds of the language may be the least successful aspect of Irish acquisition. The Irish language has a complex sound system, very different to English (Chapter 2.4) and a complex mapping of sounds to written forms (Chapter 2.6.3), which can be challenging for both pronunciation and literacy acquisition. Ideally, a pedagogical approach is needed that allows learners to get an intuitive grasp of the sounds and how they link to written forms. Unfortunately, there is a pedagogical gap as phonetic-linguistic research of Irish structure has not been carried into content development or into the classroom. Other aspects of the language structure, such as the morphology and syntax, being quite different from English, can also present challenges (Chapter 2.5, 2.6).

Overall, there is a lack of resources compared to what is available for other languages. Many of the materials available fail to engage the modern young learner, who is immersed in state-of-the-art game-based learning content, and this reduces their motivation, and contributes to a negative view of the language. In Irish-medium schools, there is a dearth of materials for the teaching of subjects other than Irish. The Government of Ireland's 20-year Strategy for the Irish Language 2010-2030 [6] has the stated aim of providing the teaching of some subjects other than Irish through Irish in all mainstream primary schools (Content and Language Integrated Learning (CLIL)). CLIL has been found to be highly effective in other countries [7] but will be difficult to implement for Irish given the lack of suitable teaching materials. This is yet another area where CALL resource development stands to make a real difference.

Of course, CALL in and of itself is not a panacea, but as a way of harnessing the resources and technologies of the Digital Plan for education, it offers unprecedented opportunities to alleviate many of the difficulties mentioned. The simple facility of having immediate access to how text is pronounced can greatly assist the learner. Bringing this to another level, interactive, multimodal, game-based platforms that are speech-enabled allow immersion in communication with virtual native speakers, offering optimal conditions for language acquisition and addressing the communicative deficits alluded to above. Recent research on speech-based multimodal educational games and virtual reality settings (described briefly below) clearly indicate increased learner engagement and motivation to persist with language learning tasks ([8, 9]. The beauty of such platforms is that the emphasis is on the spoken language and the development of communication skills, while not neglecting grammatical and written accuracy.

Importantly, these kinds of CALL environments offer a holistic immersive experience, supporting whole-language training. While a specific CALL game might have as its main target the training of some grammatical aspect of the language and exploit emerging NLP resources, the fact that learners are hearing and speaking the language as part of this game means that they are simultaneously developing pronunciation, comprehension and grammatical skills. This resonates with recent thinking on language acquisition: many modern practitioners advocate a whole-language or balanced approach whereby different aspects of the language are simultaneously acquired rather than traditional approaches which separately tackle reading, writing, comprehension and speaking [9, 10].

Digital technologies are poised to take centre-stage in the delivery of education, and Irish needs to be in step with this revolution or it will be seen as outdated. Integrating new speech and language technologies/resources into modern interactive educational platforms is vital. Ultimately, the aim is to motivate and empower learners and to recast the image of Irish as a vibrant, living language integrated in and relevant to the digital age.

14.3 Who is CALL for?

School learners: CALL has the potential to revolutionise the delivery of Irish language teaching/learning in schools. It is an exciting time in this field for Irish: the emerging speech and language technologies of the Digital Plan (Part II), allied to the linguistic knowledge and resources mentioned in Part I, offer the possibility in the not too distant future of having properly integrated intelligent CALL platforms. When such platforms deliver language learning within interactive, immersive, game-based environments there is truly an opportunity to turn around learners' experiences and engagement with the language. The potential impact for the c.700,000 learners of Irish in our school system cannot be underestimated. Platforms will need to be devised that will be pedagogically suited to all levels, including preschool.

Parents: parents doing homework with their children would benefit from carefully designed CALL platforms, e.g. pronouncing text for early readers.

Adult learners: many adult enthusiasts wish to improve their Irish in later years. Beyond these shores there are high levels of interest, as can be inferred from the c.60,000 learners from 138 different countries registered on the online course Fáilte ar Líne – Fáilte Online [11]. Irish is incredibly popular on the language app Duolingo [12] ,with recent reports stating that over 1 million people are actively learning Irish on it every week. Increasingly Irish is being offered at university level abroad and there are groups of learners througout the world, e.g. in the U.S., Canada, Britain, and a growing number in Russia and mainland Europe.

Autonomous learners: CALL is very suited to autonomous learning [12] and this is something which is particularly important for adult and foreign learners, while benefitting all learners.

Native Speakers: Although the discussion in this chapter is primarily directed at second language learners of Irish, speech and language technologies offer many direct and indirect benefits to the native speaker, as the experience in widely spoken languages shows. What varies for L1 speakers are the precise goals and levels of language targetted.

Teachers: CALL is not intended to replace teachers but should empower them by putting technology and resources at their disposal. Access to high quality, adaptive, structured CALL materials can allow a teacher to cater for different levels within a classroom, thus ensuring that every pupil is getting personalised material matching their needs as well as getting the extensive spoken language input that is otherwise difficult to deliver in a classroom. As is elaborated below (§14.5), teachers are central partners in the development and delivery of CALL, as are the curriculum developers who will ensure that CALL is a seamless part of the curriculum: the involvement of both groups must be prioritised in future CALL design and development. Teacher training will in the future encompass competence in effective CALL delivery and provide opportunities for teachers to be actively engaged in contributing to the development cycle . The Department of Education and Skills will need to be a key player in promoting this area.

Pedagogy at the heart: although there are many stakeholders, the usefulness of CALL will depend on the quality of the linguistic and pedagogical content. Stringent quality control is integral: no matter how attractively packaged, poor linguistic/pedagogical content would seriously undermine the credibility of this approach.

14.4 What's been achieved to date?

Currently Available Digital Resources:
There are very few specifically designed CALL platforms available for Irish that truly exploit the potential of speech and language technologies. There are, however, numerous online learning materials, useful NLP resources, and increasingly learner corpora that can inform future content and syllabus design.

Online language learning materials: an appetite for new multimedia frameworks for language learning is demonstrated by the growth in attractive digital materials being made available for the school-going population, e.g. Bua na Cainte [14] and Abair Liom [15] for English-medium schools; Séideán Sí [16] for Irish-medium schools; and Cód na Gaeilge [17], Féasta Focal [18] and Foghlaí Focal [19] catering to learners in Northern Ireland, as well as online materials [20] and a number of educational applications [21–24]. The website of An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG) and of Gaeloideachas are providing invaluable directories for the materials available to date [25-26].

NLP resources: very valuable NLP resources (see Chapters 4, 5) are already widely used by learners. These include spell checkers (e.g [27]), grammar checkers (e.g [28]), dictionary/terminology resources (e.g [29]). Underpinning these are corpus resources (e.g [30]), morphological analysers/generators (e.g [31]), part-of-speech taggers (e.g [32]), and treebanks and parsers (e.g [33]).

Learner Corpora: important learner and educationally-focused corpora are coming on stream (Chapter 3) that will be useful for curriculum design and assessment and to inform pedagogical content. Some of those aim to establish learners' grammatical, lexical attainment that corresponds to the levels of the Common European Framework of Reference for Languages (CEFR) [34]. Ongoing analysis of currently available educational textbook corpora [35] will allow grading in terms of complexity of lexicon and syntax.

CALL platforms: current prototypes

The Digital Plan promises to be a game changer in how we approach Irish language learning. We aspire to sophisticated intelligent CALL platforms which harness speech and language technologies, and embed phonetic and linguistic knowledge in the very heart of language pedagogy in a way that engages the young digital generation.

Adaptation of existing platforms: where existing CALL applications offer desirable features, such as attractive virtual reality scenarios or adaptive characteristics, they can be usefully repurposed. This entails taking the skeleton of a pre-existing scenario or game and using it as a vehicle for the delivery of pedagogical materials appropriate to the Irish learners' context. Digichaint is a multimodal interactive language-learning prototype game of this kind, where learners interact with a host of virtual characters in a virtual hotel setting in order to solve a problem [36]. The content was developed from scratch, and it retains the adaptive features so that learners' choices within the game determine the language level of subsequent material offered. This feature makes it suitable for pupils at different levels in a single classroom. Fáilte go TCD is a rather different prototype which allows the learner to eavesdrop on conversing groups speaking as Gaeilge in the Front Square of Trinity College [8, 37]. The conversations have been designed specifically for advanced learners and they are closely aligned to the Leaving Certificate curriculum. In both these platforms the characters speak with ABAIR synthetic voices.

New platforms targeting challenges specific to Irish: the true potential of CALL will be realised through the creation of new platforms which are designed to address the specific challenges of Irish language learning, and which embed the phonetic and the linguistic knowledge (Part I), translating them into material appropriate for the classroom. This allows for intelligent interactive platforms in a way that adapting existing CALL platforms developed for other languages can not.

One example is the game Lón don Leon, which focuses on developing an awareness of the complex system of consonantal contrasts in Irish (Chapter 2.4.2-4) as well as an intuitive grasp of how these contrasts link to the writing system, i.e. the phonics (see Chapter 2.7.3) for early literacy development. Entirely newly created content was required, including graphics, stories and music, dictated by the linguistic learning targets. This system combines prerecorded speech and songs with synthetic voices and linguistic resources of the text-to-speech system, e.g. dialect-specific letter-to-sound rules (see Chapter 8).

A further example, An Scéalaí is an intelligent CALL platform targeting the specific features of Irish structure. At its core are the synthetic voices and NLP resources. A primary feature is that it aims at holistic language learning, whereby writing and reading skills are trained in parallel with listening and speaking skills. The current learner tasks involve story writing, but from the outset the writer listens back to the spoken version (produced by synthetic voice). This prooflistening approach enables the learner to spot many of the common errors and ensures that the correspondence of the spoken and written forms is being reinforced. Text, spoken, and visual messages are also used to prompt the learner to correct errors which their ears fail to find. The pedagogical focus is currently on specific morphological and phonological features of Irish and the remit expands with the emerging resources of Part I. This platform is modular and its development will also exploit the technologies emerging in Part II as they come on stream, particularly the spoken dialogue systems described in Chapter 10.

This platform doubles up as an iCALL research tool. Learners' written and oral contributions in specific language learning tasks are retained along with their interventions in correcting their own materials. This is building a learner corpus, which is being analysed to guide linguistic priorities to sequence the learning targets in future iterations of the CALL platform. It is also yielding rich insights into the language learning process.

Communicative interactivity is key: all these platforms are interactive, but there are major language learning dividends when the platform allows a true communicative experience for the learner. The pilot platform Taidhgín (see Chapter 10) demonstrates how even the existing technology can enable the semblance of a true conversational partner that engages the learner in a chat (with a monkey)! Evaluation in schools show very high engagement on the part of the learners, even though the topics on which the monkey could converse were rather limited. The speech element here is vital: only synthetic voices can permit the open-ended utterances needed for this kind of application where the learners' input can not be predicted and thus, prerecorded. This platform is also made 'intelligent' by taking advantage of NLP resources (see Chapter 10). It is the integration of clever speech and language technologies that is the hook that engages the learners' attention and makes the learning fun.

14.5 Future Directions: building CALL capacity

CALL, is an applied field and draws on expertise and tools from very differing disciplines. Figure 14.1 below illustrates this point showing how the existing and emerging technologies of Part II and resources of Part I feed into a development where a core team interacts, not only with the technology developers but with pedagogical experts and end users. This figure maps the essential human and other resources needed in the generation of tomorrow's educational platforms for Irish.

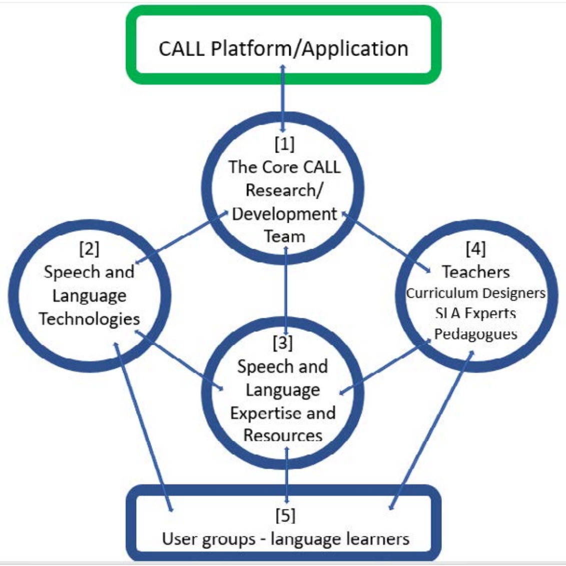

CALL Platform/Application
[1]
The Core CALL Research/ Development Team
[2]
Speech and Language Technologies
[4]
Teachers
Curriculum Designers
SLA Experts
Pedagogues
[3]
Speech and Language Expertise and Resources
[5]
User groups - language learners

Figure 14.1: Components of a CALL platform for Irish

(1) The core CALL research/development team
acts as the hub that designs, develops and tests the platforms, integrating input from the different components and working closely with teachers and learners. At the centre of the hub, the coordinator(s) design and develop content for the learner platforms, drawing together the technologies and the pedagogical content. The core team also requires interface designers and programmers to build and implement the platforms. Given the overlap of skills, the technical development team should work in parallel and in close collaboration with those developing the technology and resources.

(2) Speech and Language Technologies (Part II)
As the core technologies of Part II are developed, they will provide the capacity to develop a wide range of powerful learning platforms. Given the central importance of educational applications in the context of Irish, it is important that the priority in technology development is appropriate to the envisaged educational uses. As is clear from the above, spoken dialogue systems are particularly important. These require speech recognition, which was unavailable until recently. Note however, that with the recent advances in this field (see Chapter 9) new possibilities are opening up.

Although speech synthesis is more advanced, in order to meet the needs of the envisaged CALL applications we would aspire towards flexible voices that can be transformed to provide multiple characters that populate the games and which can be modulated to capture the expressive prosody that will make characters believable in a dialogue context (Chapter 10). However, CALL developers need not wait for fully developed technologies to be provided. As experience with Taidhgín has shown, even a primitive dialogue system, lacking ASR, can nonetheless serve as a powerful and engaging learning aid. Children's voices will also need to be prioritised for both synthesis and recognition.

High-quality machine translation would also impact on the education environment, particularly for those in Irish-medium schools. For example, in researching a school project online, learners are currently working with predominantly English materials, i.e. in an English-dominated environment, a situation which could be transformed with this technology. Embedding NLP resources and dictionaries in learning applications stands to become an integral part of future learning environments.

(3) Speech and Language Expertise and Resources (Part I)
The expertise and the growing body of linguistic resources emerging from the research in Part I are the heart of intelligent, knowledge-based CALL applications. This is essentially how we translate linguistic and phonetic research into concrete learner-appropriate form to enrich the teaching of the language. As referred to above, semantic wordnets (Chapter 4.5) embedded in an interactive application can be very effective for implicit learning of certain features of the grammar [9]. We can envisage the use of articulatory models for the teaching of Irish sound contrasts and prosody models with visual feedback (Chapter 7) can be incorporated in an interactive game that would make this abstract and neglected aspect of Irish pronunciation graspable, and fun to learn. Note that the linguistic/phonetic researchers need

to work hand in hand with the CALL core team, and, of course, with experts in pedagogy. This work will serve to illuminate aspects of Irish linguistic structure not hitherto available to teachers and learners of Irish (e.g. prosody, syntax and semantic models). Where possible, all these resources should be developed in a way that maximises their potential for exploitation in educational (and disability/access) applications.

(4) Pedagogy before technology: Teachers & Curriculum Designers

It is important to avoid a not uncommon pitfall with CALL development described by Underwood as 'letting computer whizzes first explore how the computer can do something particularly well and then design the program to take advantage of that' [38, p. 83]. The emphasis must be on pedagogy before technology, where the learning is the central goal, and technological innovation responds to pedagogical need: the educational aim and the content to deliver it must be based on a clear understanding of the learner's context, level, and outlook.

The heart of CALL is the pedagogy and it is on the quality of the content that it will ultimately be judged. It is therefore imperative that the core content developers of the CALL team collaborate closely with practitioners, i.e. teachers and curriculum designers. Expertise in the area of linguistics/second language acquisition, particularly in the Irish context, is also required. This is important so that assumptions based on English are not automatically assumed to be relevant for Irish.

(5) End-user groups

The CALL endeavor is for the learner, therefore the learner must be an integral part of every step of CALL development. The final litmus test for any platform is the degree of enthusiasm with which it is adopted by learners. It is therefore important that networks of teachers and learners be established that will collaborate with the core CALL team in enabling new developments to be used and evaluated by learners on an ongoing basis as they come on stream.

References

[1] Department of Education and Skills. (2015). *Digital Strategy for Schools 2015-2020: Enhancing Teaching, Learning and Assessment.*

[2] Mac Gréil, M. & Rhatigan, F. (2009). *The Irish language and the Irish people: Report on the attitudes towards, competence in and use of the Irish language in the Republic of Ireland in 2007-2008*. National University of Ireland Maynooth: Survey & Research Unit, Department of Sociology.

[3] Murtagh, L. (2003). *Retention and attrition of Irish as a second language: A longitudinal study of general and communicative proficiency in Irish among second level school leavers and the influence of instructional background, language use and attitude/motivation variable*. University of Groningen, Netherlands.

[4] Harris, J., Forde, P., Archer, P., Nic Fhearaile, S. & O'Gorman, M. (2006). *Irish in primary schools: Long-term national trends in achievement*. Department of Education and Science.

[5] Dunne, C. (2015). *Becoming a Teacher of Irish: The Evolution of Beliefs, Attitudes and Role Perceptions.* [M.Ed. thesis]. Trinity College, Dublin.

[6] Rialtas na hÉireann. (2010). *Stráitéis 20 Bliain don Ghaeilge 2010-2030: 20-year strategy for the Irish language 2010-2030.* Baile Átha Cliath: Oifig an tSoláthair.

[7] Coyle, D., Hood, P. & Marsh, D. (2010). *CLIL: content and language integrated learning*. Cambridge University Press.

[8] Ní Chiaráin, N. (2014). *Text-to-Speech Synthesis in Computer-Assisted Language Learning for Irish: Development and Evaluation.* [Ph.D. thesis]. Trinity College, Dublin.

[9] Ní Chiaráin, N. & Ní Chasaide, A. (2020). The Potential of Text-to-Speech Synthesis in Computer-Assisted Language Learning: A Minority Language Perspective. In A. Andujar (Ed.), *Recent Tools for Computer- and Mobile-Assisted Foreign Language Learning* (pp.149-169). Hershey, PA: IGI Global.

[10] Pressley, M. & Allington, R.L. (2014). *Reading instruction that works: the case for balanced teaching* (4th edition). New York: The Guilford Press.

[11] Fiontar agus Scoil na Gaeilge (Dublin City University) & National Institute for Digital Learning (Dublin City University). (2020). *Fáilte ar Líne – Fáilte Online.* www.failteonline.ie

[12] Duolingo. (2022). www.duolingo.com/course/ga/en/Learn-Irish

[13] Egbert, J., Hanson-Smith, E. & Chao, C. (2007). Foundations for Teaching and Learning. In J. Egbert & E. Hanson-Smith (Eds), *CALL environments: Research,*

practice, and critical issues (2nd edition, pp.19-28). Alexandria, VA: TESOL.

[14] de Brún, S. & Ní Fhatharta, M. (2014). *Bua na Cainte*. EDCO. https://buanacainte.ie/

[15] Folens. (2015). *Abair Liom.* www.folens.ie/programmes/abair-liom/products

[16] An Gúm, Department of Education and Skills, Council for the Curriculum, Examinations and Assessment (Northern Ireland) & Fóras na Gaeilge. (2015). *Séideán Sí*. www.seideansi.ie

[17] Council for the Curriculum, Examinations and Assessment (Northern Ireland). (2020). *Cód na Gaeilge.* www.ccea.org.uk/learning-resources/cod-na-gaeilge

[18] Council for the Curriculum, Examinations and Assessment (Northern Ireland). (2020). *Féasta Focal.* www.ccea.org.uk/learning-resources/feasta-focal

[19] Council for the Curriculum, Examinations and Assessment (Northern Ireland). (2020). *Foghlaí Focal.* www.ccea.org.uk/learning-resources/foghlaí-focal

[20] An Comhairle um Oideachas Gaeltachta & Gaelscolaíochta. (2020). *Léigh Leat*. www.leighleat.com

[21] Ní Arrachtáin, E. (2015). *Lingua App*. Oideas Gael. www.lingua-app.ie

[22] Coláiste Lurgan. (2015). Abair Leat *Oide.* www.play.google.com/store/apps/details?id=air.AbairleatOideMobile&hl=en

[23] Ní Ghallchobhair, E. (2011). *Enjoy Irish.* www.enjoyirish.ie/#anchor

[24] Xu, L., Uí Dhonnchadha, E. and Ward, M. (2022). Faoi Gheasa: an adaptive game for Irish language learning. ACL 2022 ComputEL Workshop.

[25] An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta. (2020). Resource Database. www.cogg.ie/bunchar-aise-anna/#?keywords=&subject=&types=&levels=&page=1&language=ga

[26] Gaeloideachas. (2020). *Gaeloideachas: guth don Oideachas lán-Ghaeilge agus Gaeltachta.* www.gaeloideachas.ie

[27] Scannell, K. (2019). *GaelSpell* (Version 5.1). www.cadhan.com/gaelspell/index-en.html

[28] Scannell, K. (2013). *An Gramadóir* (Version 0.7). www.cadhan.com/gramadoir/index-en.html

[29] Foras na Gaeilge & Gaois (Fiontar & Scoil na Gaeilge, Dublin City University). *(2022). An Bunachar Náisiúnta Téarmaíochta don Ghaeilge*. www.tearma.ie

[30] Gaois research group (Fiontar & Scoil na Gaeilge, Dublin City University). (2022). *Gaois*. www.gaois.ie

[31] Uí Dhonnchadha, E. (2015). *Natural Language Processing (NLP) Tools for Irish.* www.scss.tcd.ie/~uidhonne/irish.utf8.htm

[32] Uí Dhonnchadha, E. (2010). *Natural language processing tools: Developing a part-of-speech tagger and partial dependency parser for Irish*. Saarbrücken: Lambert Academic Publishing.

[33] Lynn, T. (2020). *The Irish Universal Dependency Treebank* (Version 2.6). www.github.com/UniversalDependencies/UD_Irish-IDT/tree/master

[34] Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR).* www.coe.int/en/web/common-european-framework-reference-languages/home

[35] Ó Meachair, M.J. (2020). *The Creation and Complexity Analysis of a Corpus of Educational Materials in Irish (EduGA).* [Ph.D. thesis]. Trinity College, Dublin.

[36] Ní Chiaráin, N. & Ní Chasaide, A. (2016). The Digichaint interactive game as a virtual learning environment for Irish. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp.330-336).

[37] Ní Chiaráin, N. & Ní Chasaide, A. (2015). Evaluating Synthetic Speech in an Irish CALL Application: Influences of predisposition and of the holistic environment. In *SLaTE: 6th Workshop on Speech and Language Technologies in Education* (pp.149-154).

[38] Underwood, J. (1984). *Linguistics, computers and the language teacher: a communicative approach.* Rowley, MA: Newbury House.

Chapter 15
Applications for Disability and Access

15.1 What is involved?

Speech and language technologies have a particularly important role to play in the lives of people with disabilities. Speech communication encompasses a chain of events [1] shown in Figure 15.1, from speaker (production) → speech signal (acoustics) → Listener (perception).

Production

- ***Cognitive/linguistic encoding:*** the speaker maps ideas and concepts to linguistic form and sends messages to the speech organs to realise them.
- ***Physiological encoding:*** motor nerves send messages to the speech organs, resulting in lung air being transformed to patterned sound by the vocal folds (voice source) and articulation (filtering) (Chapter 2.4 and 7).

Acoustic Code

- ***Acoustic coding:*** a sound wave carries the 'speech code', reflecting patterns of source and filter.

Perception

- ***Physiological-receptive Decoding:*** the acoustic signal is transformed in the listener's ear and the signal is carried to the listener's brain
- ***Cognitive-linguistic decoding:*** the essential meaningful features of the signal are extracted, and the listener reconstructs the speaker's intended message

The speech chain may be disrupted at any point due to cognitive, motor or sensory issues. Speech technology may offer a solution: for example, if one can formulate language, but cannot speak or hear, speech synthesis and recognition technologies can open up channels for communication. Written text is a widely used medium of communication. Where the ability to read or write is disrupted, e.g. for those with visual impairment, motor difficulties or dyslexia, access to the written word present barriers to communication. Here also, speech and language technologies offer ways to restore access.

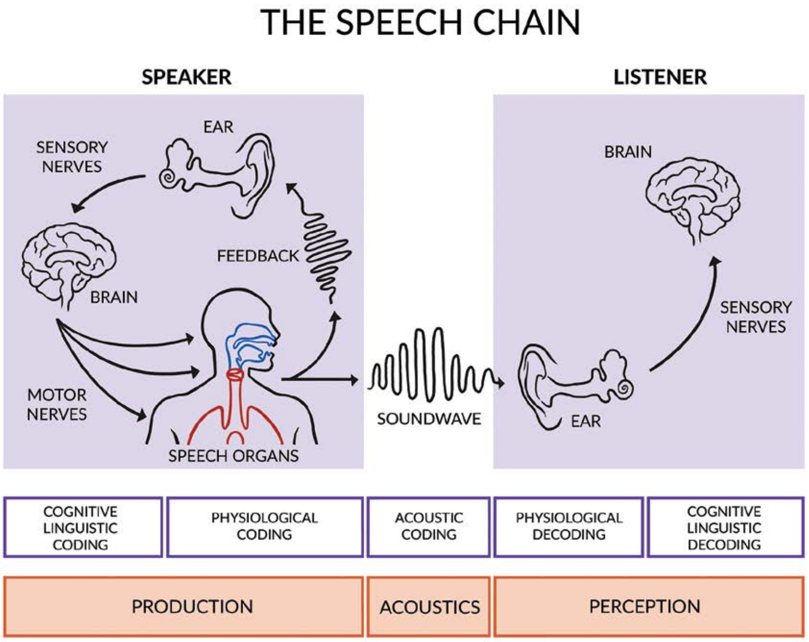


Figure 15.1: The Speech Chain (based on [1])

International figures indicate that the percentage of people living with a disability is very high – US statistics suggest 1 in 4 adults [2]; EU survey estimates are at 17.6 % of the population aged 15 and over [3]. Where disability involves a disruption to the speech chain and/or difficulties writing, reading, seeing text, it has serious consequences for every aspect of one's life. **Professionally**, many avenues are closed as the communication disrupt prevents people from doing jobs they are otherwise well qualified for. For example, in Europe it is reported that over 75% of blind and partially sighted persons of working age are unemployed [4], and similar statistics pertain in Ireland [5]. Access to **education** and more generally to information is difficult, with the consequence that education is often cut short. **Social exclusion** is a further consequence of these communication difficulties.

Speech and language technologies can enable those with disabilities to bridge the break in the communication chain and access a fuller educational, professional and social world. However, there is relatively little provision for Irish. This greatly exacerbates the situation whereby parents of children with disabilities are advised too frequently by health professionals and teachers not to pursue Irish at school, not to attend Irish-medium education: there are also cases of Irish-speaking families in the Gaeltacht being counselled not to use Irish as a home language. Such advice is misguided, based more on popular belief than on scientific evidence, as the benefits of early bilingualism have been shown repeatedly in research. There is an imperative to develop the technologies and applications for Irish that enable communication for those with disabilities, and that facilitate their inclusion in the Irish-language educational, professional and social community. The needs of those with disabilities are foregrounded in the Digital Plan, and it is recommended that *all applications are designed from the outset so that everyone can use them* – a principle articulated in the *Design* for All framework [6]. Linguistic resources (Part I of the Digital Plan) are also very important for how we provide for those with disabilities and essential to many of the applications needed.

There are many domains where technology can help. This chapter cannot attempt to cover such a broad area but presents a brief discussion on the implications for: those without (intelligible) speech (§15.2); the visually impaired (§15.3); those with literacy difficulties, such as dyslexia (§15.4); and speech and language therapy (§15.5). General recommendations are presented in §15.6.

15.2 Assistive technology applications for those with vocal disability

15.2.1 Why is it important and for whom?

Some conditions, such as cerebral palsy or autism, may result in the individual never being able to produce intelligible speech. Many illnesses also disrupt or greatly impair speech, e.g. motor neuron disease, Parkinson's disease, multiple sclerosis, stroke and laryngeal cancer.

As the well-known case of Stephen Hawking has shown, access to speech synthesis technology (Chapter 8) can enable one to remain connected and to lead a full life despite such vocal disability. For those who can type, predictive text and spelling and grammar checkers help this process. However, many with vocal disabilities also have difficulty handling a keyboard and require specialised systems (AAC devices[1]) that allow them to compose the message to be spoken by the synthetic voice. These are often tablet-based apps: the user chooses a series of pictures or symbols which are then transformed into sentences to be spoken by the synthesiser. For those with severe motor difficulties, symbols or letters can be chosen through eye gaze.

Rapid advances are occurring in this field. For those who can speak, but whose speech is unintelligible, the possibility is being investigated of training a speech recognition system to 'recognise' the speech, and then have it read out by a synthetic voice [7] – building on techniques for adapting speech recognition to a particular speaker (Chapter 9.5). For those who communicate using synthetic speech, the synthetic voice should match the user (see below and Chapter 8). CSTR[2], Edinburgh has pioneered techniques for adapting the synthetic voice, drawing on vast corpora of spoken data, collected for speech recognition. It has also pioneered personal synthesis systems for people who are losing their speech, based on recordings made of their own voice before they lose it [8].

The future possibilities are endless: complex technologies involving speech-enabled avatars will feature more and more. Systems are being piloted that

[1] Speech-based Alternative and Augmentative Communication technologies (AAC)
[2] Centre for Speech Technology Research, Edinburgh

would allow a person with locked-in syndrome (who can compose but not utter the linguistic message) to use mental commands (from the brain's motor cortex to speech organs) to control a speech synthesiser [9].

15.2.2 What is available to date?

Speech synthesis with assistive technology (AAC): speech synthesis for the three main dialects is freely available. However, there is no readily available AAC device for those who cannot operate a keyboard, but research towards this goal is ongoing, by the ABAIR[3] group at Trinity College, Dublin[4]. A very initial prototype is described in [31].

Applications using speech recognition are not available, but might become more feasible, given the advances being made in Irish speech recognition (Chapter 9).

15.2.3 Future needs

AAC applications with speech synthesis: for those who can operate a keyboard, text prediction is needed to speed up text input to the synthesiser. For those who cannot operate the keyboard, developing a fully functional Irish AAC system is an urgent priority. This is not a simple matter of interfacing an existing (English) system to an Irish synthesiser, or a matter of translating the words for the pictures/symbols in an English AAC system to Irish. Basic linguistic research is needed for a system that generates speech from concept (pictures/ symbols) – respecting the linguistic structure of Irish (which is quite different from English). Given the bilingual context, the system needs to be designed to allow sentences to be generated in both Irish and English in an optimal way for the user. The synthetic voice also needs to be 'able' to speak in Irish and English.

The synthetic voice and user identity: our voice is intimately linked to our identity, and it is important that the synthetic voice respects the identity of the disabled user. At the very least, it must be appropriate, in terms of the user's dialect, gender, and age: catering for children's voices is of particular importance. Furthermore, when the synthetic voice is your voice, control of voice quality is very important in two ways (see Chapters 7 & 8):

- ***Fine tuning the synthesis baseline voice quality*** is desirable to match the user's own sense of identity and personality (soft voice, strident voice etc.);
- ***Enabling the expression of emotion and attitude:*** when conversing, we vary the voice quality continuously, (voice prosody) to sound angry, friendly, sad etc. Having speech output but being unable to express emotion in the voice can be very demoralising for the disabled user.

Applications using speech recognition: as speech recognition comes on stream, applications might be feasible which would fine tune the system to the individual's voice, in a way that would render relatively unintelligible speech as text or synthetic speech output.

This research and development will need collaboration with end users, their carers, therapists and with disability organisations.

15.3 Applications for the visually impaired

15.3.1 Why is it important and for whom?

Not having access to the written word has major consequences for education, work, social interactions etc. According to the 2016 census, there are 54,810 people in Ireland who are blind or visually impaired, and 4,701 are children of school age (5-18 years) [5]. Children in Gaeltacht and Irish-medium schools face particular difficulties, and often considerable pressure to switch to English to pursue their education.

Screenreaders with synthetic speech output are an essential application for those with little or no sight, allowing them to hear text on computer and to navigate the computer's software by ear. Audiobooks are also helpful and are especially needed for schoolchildren's textbooks. Dictation systems with speech recognition are a further key application to facilitate text composition from voice.

15.3.2 What is available to date?

Screenreader: Irish synthetic speech is now available as a plug-in with the free, opensource NVDA screenreader[5] [10]. Its development in the ABAIR group at Trinity College Dublin was prompted by urgent requests from parents of visually impaired children. It was also supported by the NCBI[6] and COGG[7] and was tested with school pupils by teachers of the visually impaired. The user choses the dialect, controls the

[3] On the ABAIR website: www.abair.ie . ABAIR is supported by the Department of Tourism, Culture, Arts, Gaeltacht, Sport & Media.
[4] Partial additional funding comes from An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG).

speed of speech output, and can highlight and magnify text as it is read out. It also works with the Liblouis Braille system to provide simultaneous Irish speech and Braille output, as illustrated in Figure 15.2.

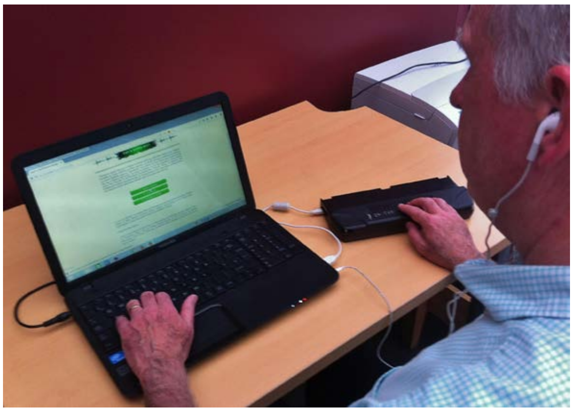


Figure 15.2: The NVDA screenreader with ABAIR plug-in in use with Liblouis Braille system

DAISYBooks [11]: multimedia versions of school textbooks featuring the ABAIR synthetic voices are available to visually impaired children through *Childvision Ireland*.

Dictation systems with speech recognition are not yet available.

15.3.3 Future Needs

Dictation systems are needed recognising a wide variety of speakers, dialects and, most importantly, children's voices[8].

Screenreaders: although the NVDA screenreader with the Irish plugin allows Irish speech and Braille outputs, interfaces for other widely used devices are also needed.

'Design for all' applications: for those with partial vision, many general-purpose applications are potentially very useful. These include web-reading facilities with highlighting and magnification of text, educational language learning games, literacy aids etc. (Chapters 8, 13, 14). Thus, in all mainstream application development, the needs of the partially sighted – and of all with disabilities – must be considered.

For complex disabilities: some people with visual impairments have complex disabilities and therefore, the kinds of interfaces and peripherals discussed in §15.2 will also be needed.

The involvement and collaboration of user groups, specialist teachers, therapists and organisations for those with visual disabilities is essential.

15.4 Applications and resources for those with literacy challenges, such as dyslexia

15.4.1 Why is it important and for whom?

It is estimated that 10% of the population of Ireland has dyslexia [12]. As there are over one million students in full-time education in Ireland [13], this would put the number of students with dyslexia in Ireland at over 100,000. While some of these students get exemptions from studying Irish, many choose to study the national language. Pupils in the Gaeltacht and many outside pursue their entire school education through the medium of Irish. (International research indicates major benefits of early bilingualism.) In the context of both equality in education and linguistic equality, it is of paramount importance that all pupils have the option to learn Irish/be educated through Irish and have access to the necessary supports.

Technology applications are widely used to aid literacy acquisition in English and are essential in allowing those with literacy difficulties to participate fully in education and in professional spheres. Dictation systems using speech recognition are widely used for composing text. Spell and grammar checkers are an invaluable resource. Speech synthesis is used to read text aloud. It supports reading acquisition, as it reinforces the connection between spoken and written forms and is useful for aural correction of text (proof-listening).

Linguistic research (Part I) is particularly important for the development of resources for literacy screening, assessment and learning support. These must be grounded in:

(i) a knowledge of the sound structure and writing system of Irish: these structural features greatly influence the ease and speed of literacy acquisition (§2.7.3). Irish has a rich system of consonants, which

[5] The plugin for Irish synthetic voices for the NVDA screenreader can be downloaded from the ABAIR website at www.abair.ie/nvda. ABAIR is supported by the Department of Tourism, Culture, Art, Gaeltacht, Sport & Media.
[6] National Council for the Blind of Ireland
[7] An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta
[8] The Irish language DaisyBooks involved an ABAIR collaboration with ChildVision Ireland and was supported by An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG).

is different from that of English (§2.4). Likewise, the way sounds map to the written forms (the phonics system) differs greatly between the two languages. The Irish system, though opaque, is quite regular (compared to the English) and therefore well suited to a phonics-based approach to literacy instruction. Nonetheless, current practice is based mainly on whole-word recognition [14], which is far from optimal (see §2.7.3). This impacts all learners but particularly those with dyslexia. The differentiation of Irish consonants is at the heart of Irish phonics: therefore, instruction in phonics must build on phonological awareness of the native sound distinctions.

(ii) an understanding of the process of Irish literacy acquisition and of how this is influenced by the bilingual context.

15.4.2 What is available to date?

Dictation systems are not available, but speech recognition is being developed (Chapter 9).

Spell- and grammar checkers are widely used and can be purchased from Cruinneog.[9]

Speech synthesis is freely available, along with resources such as a webreader[10] and is increasingly featuring in pedagogical and other applications (see below and Chapter 8).

Tools and resources for literacy assessment: there is a major gap in provision for Irish. Screening and assessment materials are all based on English – and entirely unsuitable for Irish [15]. Educational psychologists, who assess Irish-speaking pupils, and specialist teachers simply do not have the tools they need [16].

Basic research on literacy acquisition initial research in this area has been carried out and data collected in Gaeltacht and Irish-medium schools [17-21], but much more is required.

Applications for literacy support and training: the above research is guiding technology applications being developed in Trinity College's ABAIR group, exploiting speech synthesis and resources to support literacy acquisition[11] (Chapter 14). Note that these are intended for the general learner but should be especially helpful to those with literacy difficulties and may serve as platforms for future targeted learning support applications. For early learners, a game-based interactive platform, Lón don Leon [22], aims at training phonological awareness and early literacy skills, using songs, stories, pictures and activities, created to highlight the sound and phonic structure of Irish. For more advanced learners who can use a keyboard, the learning platform An Scéalaí aims to train writing and reading alongside speaking and listening skills [23]. Learners' written work is read aloud by the synthetic voice, reinforcing the connection of spoken and written forms and allowing learners to proof-listen their compositions. Spell- and grammar checkers and an online dictionary are integrated. Corrective feedback is provided in text and spoken (synthetic) forms. Learner data is collected for research and to guide future development.

15.4.3 Future needs

Dictation systems for a wide range of speakers, dialects, and especially, children's voices.

Synthesis based systems: with voices for different speakers, dialects and children.

Spell and grammar checkers will need updating as the Digital Plan progresses.

Research on Irish literacy development and dyslexia: Extensive research on literacy acquisition in Gaeltacht schools, Gaelscoileanna and in English-medium schools is needed to provide (i) wide-ranging normative data for screening and assessment tools; and (ii) the foundation for learning-support applications. This research will benefit all learners.

Online tools for literacy screening and assessment: a full suite of online tools is needed, appropriate to the sound structure and phonic regularities of Irish. Screening tests entail the testing of phonological awareness (ability to discriminate and manipulate the sound contrasts of the language) and online digital materials with native phonological contrasts are needed to ensure tests are not invalidated if administered by practitioners who do not have native-like contrasts. Materials for tests also need to be dialect-sensitive.

Literacy learning-support applications: programmes are needed to support both learners with literacy difficulties and their teachers. These must be research-based, respect the linguistic structures of Irish and will need

[9] Spell checkers and grammar checkers can be purchased at: www.cruinneog.com
[10] A webreader with Irish multi-dialect speech output can be downloaded at https://www.abair.tcd.ie/products/webreader/index.htm
[11] These applications were also supported by the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media and An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (COGG)

continuous refinement as user data and research findings become available. Integrating different technologies and resources (speech synthesis, recognition, spell- and grammar checkers, dictionaries etc.) will allow for increasingly powerful tools incorporating native models of the language. Close cooperation is needed with all language-learning application development.

Specialist training: professionals in the field, educational psychologists and support teachers will need opportunities for specialist training (see §15.6)

15.5 Applications and resources for speech and language therapy

15.5.1 Why is it important and for whom?

Speech and language therapists work with a wide variety of people with communication impairments and are often the link between the research groups developing resources and applications and the potential end-users. All aspects of the Digital Plan are relevant to therapists working with Irish speakers and their input in the process of development is essential to ensure its full impact.

At the most fundamental level, speech and language therapy provision for Irish speakers requires protocols and assessment tools based on normative data, as well as intervention programmes for specific therapies, designed for Irish. Above all, speech therapists are needed who have a full understanding of the Irish linguistic system, how it is acquired and an appreciation of influences of the bilingual context.

15.5.2 What is available to date?

Professional training for speech and language therapy in Ireland is conducted exclusively through English. There is no specific training in the linguistics of Irish, its acquisition and other relevant fields for the therapists who provide for the specific needs of Irish speaking clients.

A number of studies have been carried out in the National Universities of Cork and Galway on the assessment of Irish speakers and aimed at adapting specific existing English-language assessment tools to Irish [24-29]. There is a group of Galway-based therapists, NEART, with an interest in various aspects of language acquisition among Irish-English bilinguals. Nonetheless, there remains a major gap in provision and a lack of resources for Irish.

Technology is not currently used in diagnosis, assessment or therapy interventions for Irish, and assistive technologies are not available either, other than the screenreader for the visually impaired, described above.

15.5.3 Future needs

Each section of the Digital Plan has a potential impact. The linguistic research and resource development of Part I is critical, extending our understanding of the structure of Irish and providing the data on Irish language and literacy acquisition in the bilingual context – needed to establish the developmental norms for assessment tools in Irish therapy.

The core technologies and the applications of Parts II and III respectively have the potential to impact many domains of therapy, from providing assistive communication technologies to the provision of assessment tools and applications for therapy intervention. Many applications that would be envisaged for general Irish-language education could readily be adapted to support therapy interventions for specific speech and language difficulties. For example, models of Irish articulatory and prosodic patterns (Chapter 7) with visualisation and feedback should greatly reinforce traditional therapy. Similarly, applications aimed at training phonological awareness and early literacy (§15.4 above) could also be adapted for clinical use.

In order for the research, resources and applications of the Digital Plan to have an impact on Irish-language speech therapy provision, **specialist training in the relevant areas (see below) is essential.** This would increase the professional effectiveness and confidence of Irish-language therapists [30]. Such specially trained therapists are needed as collaborators in all aspects of the relevant research, from basic research on language and literacy acquisition, to the design of specific tools for diagnosis, assessment, applications for specific therapy intervention, the testing and dissemination of tools and applications.

15.6 Recommendations

- **Research and normative data collection,** for the development of assessment tools and of teaching support and therapy applications.
- **Applications, tools and resource development** that are research-based and designed for Irish. These should include assistive communication devices (AAC); interfaces (e.g. for screenreaders); online tools for screening and assessment of

speech, language and literacy; applications for learning support and therapy interventions.

- **'Design for All' [6] principles** in every aspect **of development.**
  - ***All applications*** are designed to be maximally usable by those with disabilities;
  - ***Core technologies*** are built to suit all end users, e.g. speech recognition and synthesis systems to suit a wide range of speakers, dialects and especially, children.
  - ***Impact statement*** should be included in all funding applications, to clarify how the research impacts those with disabilities.
- **Disability Networks** to link research and disability groups, advise on research priorities and collaborate in development. Networks might include:
  - ***A broad-based Disability Advisory Network:*** communicating the needs of those with disabilities to research groups, and the potential of the Digital Plan to those with disabilities. This network might include speech therapists, educational psychologists, support teachers, carers, disability organisations, State departments, individuals with disabilities and those who advocate for them.
  - Specific networks – user groups and professionals for specific disabilities are needed to collaborate in the design, evaluation and dissemination of particular applications.
  - ***Professional Irish-language Expert Groups – Speech Therapists/Educational Psychologists/ Support Teachers*** should be officially established, ensuring a nationwide outreach that extends to all Gaeltacht areas and to Irish speaking families and learners outside the Gaeltacht.

- **Specialist Training and Support:** professional expert groups also need to be supported and equipped as potential research partners. Additional professional training programmes should be offered, covering essential areas of knowledge for all involved in Irish speech, language, literacy and communication interventions: Irish linguistic structure and phonics (and how these differ from English); Irish speech, language and literacy acquisition and how they are influenced by the bilingual context; bilingualism research on the cognitive benefits for the child, dyslexia etc. Training in the use of all emerging resources, technologies and applications should be included.

- **Public access to information, expert advice and support:** parents, carers and individuals with disabilities need access to well-informed guidance and support. The current situation where advice to parents is often misguided is entirely unsatisfactory – particularly when it comes to advice on switching the home language, removing the child from Irish language instruction etc. A central online helpline is recommended, connecting to trained professional expert groups and to academic researchers.

References

[1] Denes, P.B. & Pinson, E. N. (1993). *The Speech Chain: The Physics and Biology of Spoken Language*. W. H. Freeman & Co Ltd.

[2] U.S. Centers for Disease Control and Prevention. (2018). https://www.disabled-world.com/disability/statistics/1in4.php

[3] European Commission. (2022). *Disability and unemployment*. https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Disability_statistics_-_need_for_assistance#Disabilities_and_the_labour_market

[4] EBU. (2022). *Euroblind.* http://www.euroblind.org/about-blindness-and-partial-sight/facts-and-figures

[5] National Council for the Blind, Ireland (NCBI). https://www.ncbi.ie/facts-about-sightloss/

[6] Stephanidis, C. (Ed.) (2001). *User Interfaces for All: Concepts, Methods and Tools*. Lawrence Erlbaum Associates.

[7] Feifei, X., Barker, J. & Christensen, H. (2019). Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing.*

[8] Yamagishi, J., Veaux, C., King, R. & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology 33*(1), 1-5.

[9] Guenther, F. H. (2015). The neural control of speech: From computational modeling to neural prosthesis. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

[10] McGuirk, R. (2015). *Exploration of the Use of Irish Language Synthesis with a Screen Reader in the Teaching of Irish to Pupils with Vision Impairment.* [M.Phil. thesis]. Trinity College, Dublin.

[11] Daisybooks. http://www.daisy.org/daisypedia/daisy-digital-talking-book

[12] Dyslexia Association of Ireland. https://www.

dyslexia.ie/

[13] Department of Education & Science. *Key Statistics.* http://www.education.ie/en/Publications/Statistics/Key-Statistics/

[14] Stenson, N. & Hickey, T. (2018). *Understanding Irish Spelling*. An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta.

[15] Nic Aindriú, S., Ó Duibhir, P. & Travers, J. (2021). A Survey of Assessment and Additional Teaching Support in Irish Immersion Education. *Languages 6*(2), 62. [16] Barnes, E. (2017). *Dyslexia Assessment and Reading Interventions for Pupils in Irish- Medium Education: Insights into current practice and considerations for improvement*. [M.Phil. thesis]. Trinity College, Dublin.

[17] Barnes, E., Ní Chasaide, A. & Ní Chiaráin, N. (2018). The design and pre-testing of literacy and cognitive tasks in Irish and English. In *Proceedings of Literacy Association of Ireland 42nd International Conference.*

[18] Barnes, E., Ní Chasaide, A. & Ní Chiaráin, N. (2021). Bilingual phonological awareness: when interdependence becomes interference. In *Proceedings of the 15th Congress of the International Association for the Study of Child Language.*

[19] Parsons, C. E. & Lyddy, F. (2016). A longitudinal study of early reading development in two languages: comparing literacy outcomes in Irish immersion, English medium and Gaeltacht schools. *International Journal of Bilingual Education and Bilingualism 19*(5), 511-529.

[20] Murphy, D. & Travers, J. (2012). Including young bilingual learners in the assessment process: A study of appropriate early literacy assessment utilising both languages of children in a Gaelscoil. *Special and inclusive education: A research perspective*, 167-185.

[21] Barnes, E. (2021). *Predicting dual-language literacy attainment in Irish-English bilinguals: language-specific and language-universal contributions.* [Ph.D. thesis]. Trinity College, Dublin.

[22] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy, A., Barnes, E. & Gobl, C. (2019). Leveraging phonetic and speech research for Irish language revitalisation and maintenance. In *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, (pp. 994-998).

[23] Ní Chiaráin, N. & Ní Chasaide, A. (2019). An Scéalaí: autonomous learners harnessing speech and language technologies. In *SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*.

[24] O'Malley, M.P. & Antonijevic, S. (2020). Adapting MAIN to Irish (Gaeilge). *ZAS Papers in Linguistics 64*, 127-138.

[25] O'Toole, C. & Fletcher, P. (2012). Profiling vocabulary acquisition in Irish. *Journal of Child Language 39*(1), 205-220.

[26] O'Toole, C. & Fletcher, P. (2008). Developing assessment tools for bilingual and minority language acquisition. *Journal of Clinical Speech and Language Studies 16*, 12-27.

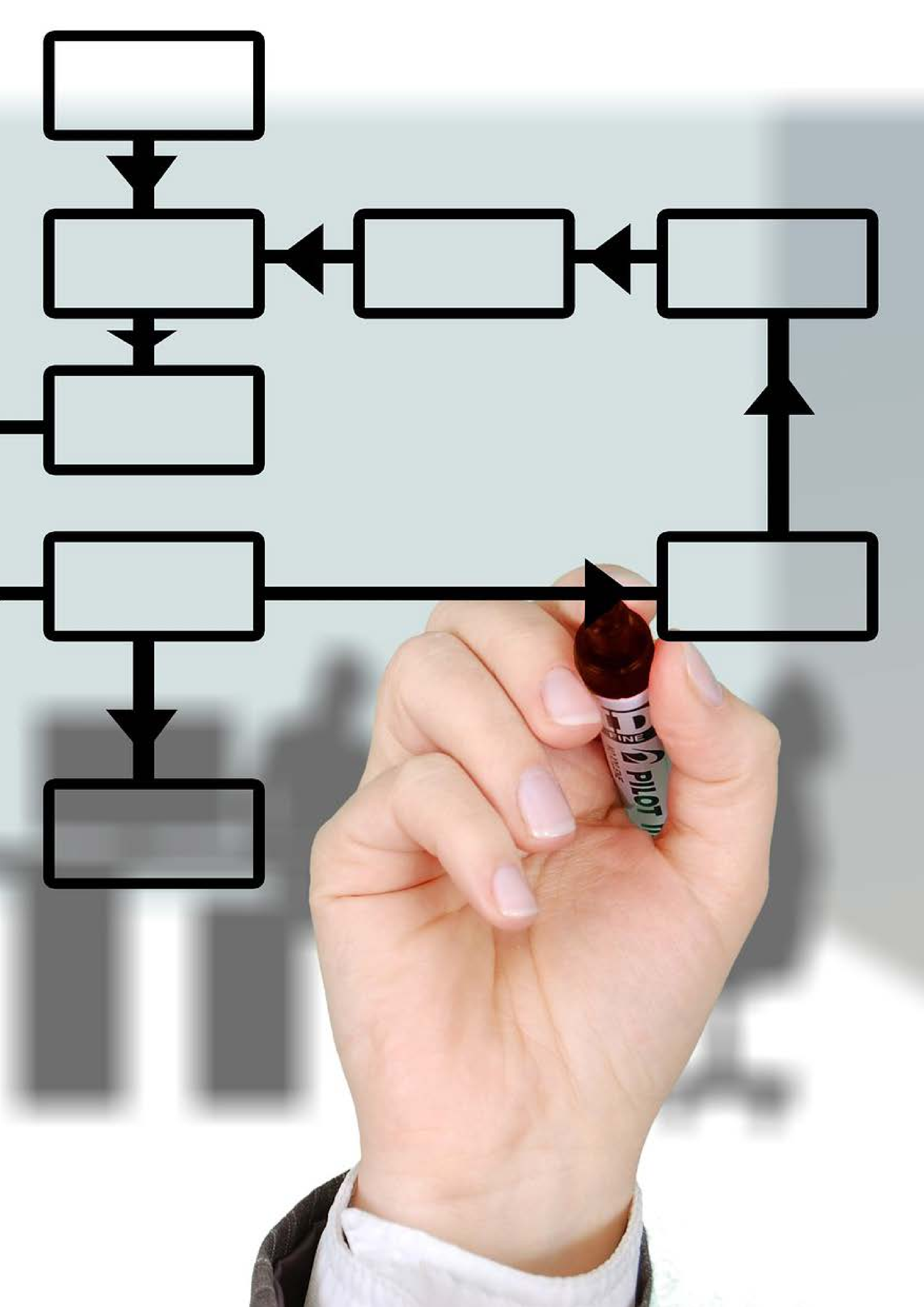
[27] O'Toole, C. (2013). Using parent report to assess bilingual vocabulary acquisition: A model from Irish. In V.C.M. Gathercole (Ed.), *Solutions for the Assessment of Bilinguals.* Bristol: Multilingual Matters.

[28] O'Toole, C. &Hickey, T.M. (2013). Diagnosing language impairment in bilinguals: Professional experience and perception. *Child Language Teaching & Therapy 29*, 91-109.

[29] Antonijevic, S., Durham, R., & Ní Chonghaile, Í. (2017). Language performance of sequential bilinguals on an Irish and English sentence repetition task. *Linguistic Approaches to Bilingualism 7*(3-4), 359-393.

[30] Harnett, S. & O'Toole, C. (2010). Speech and language therapists training and confidence. *Journal of Clinical Speech and Language Studies 17*, 57-73.

[31] Barnes, E., Morrin, O., Ní Chasaide, A., Cummins, J., Berthelsen, H., Murphy, A., Nic Corcráin, M., O' Neill, C., Gobl, C. & Ní Chiaráin, N. (2022). AAC don Ghaeilge: Prototype Development of Speech-Generating Assistive Technology for Irish. In *Proceedings of the CLTW 4 @ LREC2022* (pp. 127-132).



PILOT

Conclusions

Chapter 16
Considerations for Implementation

16.1 The Digital Plan: an integrative, long-term process

Although presented as a series of individual chapters/areas, the Plan should not be viewed as a checklist of discrete items to be delivered, but as the establishment of a broad programme with interlocking parts that complement each other to deliver much more than the sum of the parts. Research aimed at one specific technology or resource will necessarily embrace others, and as they come on stream, they will be fused in new ways, not even dreamed of today.

This Plan targets much more than the development of specific items of technology, and is intended as an initial roadmap for a long term, open-ended process. The bigger picture is the establishment of an Irish-language technology sector, intimately linked to the Irish language community it serves. In this chapter and in the Introduction (Chapter 1) consideration is given to activities and development approaches that should ensure that the Digital Plan is maximally beneficial for the language community and for the preservation of the language.

A first priority is to put in place the local infrastructure, including the human resources, for continuous development of Irish speech and language technology, grounded in a scientific understanding of Irish linguistic structure and intimately linked and responsive to the local needs. Secondly, to facilitate the full development cycle, from *research* ➔ *core technology* ➔ *applications* ➔ *dissemination*, interdisciplinary teams are needed, where different streams of research and development proceed in parallel with interaction among them. Thirdly, the need for public partnership in every aspect of the Plan is emphasised.

16.2 Who is it for?

The Digital Plan is for everyone. It aims to serve all involved at any level with the Irish language, be they native speakers from the Gaeltacht, speakers from outside the Gaeltacht, students who learn Irish at school as a second language, their teachers and parents who assist at home as well as learners at different levels here and abroad, immigrants – all those who, for one reason or another, have an interest in the language. A particular focus is on those with disabilities who, for the want of appropriate technology supports, have often been particularly disadvantaged within the Irish language community.

16.3 Impact of the Plan

Strengthening Irish in the digital space will tap into the power of modern technology to build and maintain language communities, connecting otherwise disparate and geographically separated sectors of *Pobal na Gaeilge*. The digital availability of Irish should normalise its use in ever-expanding spheres of activity. This will make it possible for young *digital natives* to make the language integral to their day-to-day social lives, bridging the current divide between their linguistic heritage and their daily interaction with digital technology.

With appropriate investment, there is the potential to transform the teaching/learning of Irish. In this, different aspects of the plan will contribute. Core technologies open the door to new teaching platforms, e.g. speech-enabled multimodal educational games, where learners might interact with virtual characters. Such applications offer a fun way to develop communicative competence in the language, to increase learners' motivation and exposure to (virtual) native speakers. The content of such games should deploy the linguistic knowledge and resources of part I. Likewise, technology-based learning research should guide how curricula and materials for teaching, learning and assessment are constructed – for example by providing a detailed, empirically-based picture of the stages which the learner passes through in moving from novice to proficient speaker. It should also enable tools for assessment and remediation, to support learners with difficulties and their teachers.

In a variety of ways, by empowering the Irish language community, by impacting positively on the learning experience of students in the education system both inside and outside of the Gaeltacht, the Digital Plan will further the goal of *An Straitéis 20 Bliain don Ghaeilge* [1], of creating a truly bilingual society.

16.4 Pobal na Gaeilge as active partners

The ultimate impact of the plan will depend on the level to which the Irish language community engages with it. The development model cannot be the same as for languages like English, where a product is first developed and then subsequently marketed to its potential public, and where commercial considerations are paramount. For Irish, as for other endangered languages, research and development priorities must be set according to the most urgent needs of the language community. Consequently, although research groups lead this work, close interaction with the language community – the technology end users – is vital from the outset, not as passive recipients of emerging applications, but as active partners, involved, for example in:

- **Identifying research priorities,** to ensure they address urgent public needs.
- **Design of applications:** input at the design stage helps ensure technologies are effective and relevant to the user.
- **Direct participation in development:** in many areas direct input from individuals and specific groups is needed. (see discussion of crowdsourcing and networks in §16.9).
- **Testing:** core technologies and applications need ongoing testing by end-users to ensure they are fit-for-purpose. Testing specific applications can provide an ideal way to also test embedded core technologies (e.g. testing the quality of synthetic voices as part of the evaluation of an educational game [2]).
- **Dissemination:** ongoing public outreach is needed to ensure that the Plan's outputs are widely used.

There are further suggestions in the following sections on how the active engagement of the Irish language community can be fostered.

16.5 Research centres of excellence: research infrastructure

Research Centres of Excellence will provide the hub for research and development of the Digital Plan. They require:

- **Interdisciplinary research groups** combining technical (engineering and computing) skills with a high level of competence in Irish and Irish linguistics. This will entail building on existing strengths and extending the scope and scale of research.
- **Continuity of funding**, essential for long term development.
- **Translating research outputs into public applications.** This is crucial and will be facilitated by having research groups engage in parallel basic and applied research. This will also require ongoing technical support, as discussed in §16.8 below.

16.6 Educating interdisciplinary researchers: the human infrastructure

The human resource infrastructure is the single most essential prerequisite of the Plan. In the current research groups, the greatest roadblock to date has been the lack of researchers with the necessary technical and Irish language/linguistics skills. Whereas in other (English-based) projects researchers can readily be recruited from across the globe, for the Digital plan, this is not an option. Key investment in appropriate interdisciplinary educational programmes is needed to train a pool of researchers, equipped with a knowledge of Irish, an understanding of Irish linguistic structure and specialist training in speech engineering or natural language processing. For the future generation of research leaders, the following is needed:

- **Interdisciplinary undergraduate and postgraduate taught programmes** that provide the skill profile needed. Existing interdisciplinary programmes that meet these requirements to varying degrees are the undergraduate degree in Computer Science, Linguistics and Language (Irish) and the taught Master's in Speech and Language Processing at Trinity College, Dublin. They have been a major source to date of skilled researchers to work on Irish speech and language technology. Such programmes should be strengthened and extended to align them to the research objectives of the Digital Plan and to provide graduates with the necessary range of technical and linguistic skills. (See also suggestions in §16.7 regarding the potential for partial Gaeltacht-based delivery of such courses.)

- **Studentships** to encourage native speakers and competent non-native speakers to pursue such multidisciplinary programmes.
- **PhD studentships and postdoctoral funding**, for researchers working in groups that are delivering on areas of the Digital Plan. This would create opportunities for graduates of taught programmes to pursue this career option, enabling expansion of the Plan's scope and providing the research leaders of tomorrow. Where specialists without Irish are employed on these projects, they should be given opportunities to avail of taught programmes to learn Irish and become familiar with Irish linguistic structure.
- **Entrepreneurship training**. This is currently promoted in third-level institutions and should be included in undergraduate and postgraduate programmes.

16.7 The Digital Plan and the Gaeltacht

Considerations should be given to how the Plan links to and impacts the Gaeltacht native-speaker community, as the repository of Irish as a living, but fragile community language. The fact that the digital world has to date largely excluded Irish means the social and other online activities of the younger, digital generation are increasingly divorced from their language. The points raised in §16.4 - §16.6 apply here, but additionally, the involvement of the Gaeltacht native speakers should be fostered, as they are needed to provide:

Expert models of the language: they are the 'expert' speakers for the extensive speech corpora in the various dialects, fundamental for linguistic research and the development of speech technologies. Recordings in the field and targeted crowdsourcing will be needed (see §16.9).

Future researchers: it is hoped that native speakers will feature strongly as skilled researchers in the long-term development of the Digital Plan. Steps to encourage their participation might include:

- **Engaging the young through ground-level outreach activities,** for example school visits by researchers; workshops for children and adolescents, where they can play with the emerging technologies; transition year visits to research labs – activities that raise awareness of Irish speech and language technology as an area for future study and employment.
- **Active recruitment to interdisciplinary, taught programmes at under-and postgraduate levels** mentioned in §16.6. As mentioned there, studentships would make such third level programmes affordable and attractive.
- **Partial location of taught programmes in Gaeltacht:** the feasibility of having some parts of the taught programmes delivered in Gaeltacht locations, as mentioned in §16.6 above, should be explored. This would require careful planning and could be facilitated by online teaching.

Future entrepreneurs: speech and language technologies lend themselves to downstream commercialisation of specific applications (see §16.10). Commercial initiatives focusing on the language would find a natural home in the Gaeltacht, helping to provide skilled employment for skilled graduates in their local community. In the case of a campus company associated with a research centre (University or other), co–location should be considered: the commercial activity might be situated in a Gaeltacht digital hub, while the research group would ensure ongoing connection to the developments in this rapidly evolving field. This is a consideration for the longer term but is worth planning for.

Encouraging co-location of educational and commercial activities arising from the Digital Plan is important to ensure that new opportunities that are created for native speakers do not result in removing them from their communities, further depleting the already fragile Gaeltacht Irish-speaking community. One of the advantages of digital technology is that it can enable the younger Gaeltacht generation to pursue interesting, challenging careers while remaining within their language community, contributing to its future.

This particular focus on Gaeltacht initiatives should not be interpreted to mean that initiatives outside the Gaeltacht should not also be actively fostered. An increasingly wide engagement of all sectors of the Irish language community is envisaged, wherever it might be.

16.8 Technical infrastructure: ongoing system supports

Given the emphasis on delivery of the entire chain from basic research to end-user applications, two-levels of technical support are recommended.

Local: software support is needed within research groups, to translate pilot, proof-of-concept platforms into robust, user-ready systems. Interfaces are also needed to integrate technologies with existing external systems. Debugging and extensive testing is required for all software. Continuous updating is needed to ensure user applications keep up with routine changes in technology, such as updated operating systems and new devices.

Central: a more centrally located technical support unit may also be needed in time, to ensure that technologies, emerging from different groups, when ready for public use, are widely disseminated and supported in schools, offices and in the home. Such a central unit should play an active role in promoting the use of newly emerging technologies as well as those already available.

16.9 Building synergies with the public and with key players

To achieve the greatest impact and connection with the public, interaction with specific networks and interest groups will be needed. The quality of much of the research and development to date speaks to the level of cooperation with such networks (e.g. teachers, translators, the visually impaired). There are many networks, of which the following are particularly relevant:

Irish Language Networks: Connecting research groups with the numerous **Irish language organisations** and **community groups** in Ireland, as well as **interest groups in other countries**, is important. The vibrant Irish-language **media sector** provides a ready-made network that reaches deep into the Irish language world: it is a potential key partner that should both directly benefit from the plan's outputs and help ensure the public is updated on all developments. Social media networks can reach those who are not part of Irish language groupings, but who would welcome the opportunity to become involved.

The Government: The Digital Plan is a national plan. It stands to benefit all Government Departments, and has particularly obvious implications in the Departments of Education, Health, Equality, Disability, Inclusion and Youth and the Department of Further and Higher Education, Research, Innovation and Science. Although initiated by the Department of Tourism, Culture, Arts, Gaeltacht Sport and Media, the Plan requires cross-departmental, cross-party support. Large-scale funding is needed, and it is important that this does not take away from funding currently available to Irish language and Gaeltacht communities and organisations, which are working to maintain the language.

The Digital Plan provides one important mechanism whereby the Irish Government can substantiate the constitutional position of Irish as the first language of the State. It is also a step towards implementing the provisions of the *Official Languages Acts* (2003 and 2021) [3] and associated recommendations of the *Digital Strategy for the Irish Language* [1] and *The Action Plan for the Irish Language, 2018-2022* [4]. Among other things it should assist in providing bilingual services and ensure that online communication with State bodies, whether text or speech-based, can be conducted through Irish. It also enables the State, in keeping with the *Disability Act 2005* [5], to fulfil its obligations to provide access to services to those with disabilities 'where practicable and appropriate'.

Other than through funding and promoting the Digital Plan, Government Departments can intervene in many practical ways to support specific areas. For example, they can ensure that all by-products of the translation process (e.g. translation memories, terminology bases and glossaries) are returned by the translation services as part of the final end-product delivery, and fed to research groups to improve the performance of machine translation systems. When specific applications, such as an Irish web-reader are developed, they should ensure they are deployed across Government Departments. Active, coordinated intervention from Government departments is also needed for the public dissemination of systems developed, so that all emerging technology resources are provided to businesses, schools, to State and Semi-State bodies, as well as to the individual citizen.

Crowdsourcing and online networks: crowdsourcing, the use of online public networks, will contribute to specific research-related activities. Crowdsourcing can be an ideal strategy to get public support for large-scale, manual, time-consuming tasks such the collection and/or annotation of certain kinds of data. However, for a minority language like Irish, where a large proportion of likely contributors are likely to be learners with varying degrees of proficiency in the language, crowdsourcing tasks must be carefully chosen and care must be taken to ensure the quality of the data obtained. Varying the crowdsourcing methodology, to allow targeting of subsections of the 'crowd' or active monitoring of the data are strategies that may be needed: this will be dictated by the type and quality of data sought. Thus, crowdsourcing might be:

i. **Open**, where a high level of linguistic or technical expertise is not needed. This was successfully used to extend the databases of Dúchas.ie [6] and Meitheal Logainm.ie [7].

ii. **Monitored**, so that tight editorial control is provided by an expert group, as was the case in the localisation of services, such as Gmail [8] and WhatsApp [9].

iii. **Targeted** to differing sectors of the 'crowd'. For example, the MíleGlór crowdsourcing of corpora for speech recognition [10] differentiates input from native speakers of different dialects, fluent non-native speakers, learners. Preselection and training of the 'crowd', i.e. potential contributors, is a strategy that has been successfully used to extend the Welsh Wikipedia entries.

Education Networks: There is a particular focus in the Digital Plan on its potential impact on Irish language education and on education through the medium of Irish. This is relevant to all levels, from pre-school to 3rd level and to informal educational settings, in Ireland and abroad. Collaboration with educational organisations, teaching bodies and networks is vital. The input of content developers, learner groups and teachers is central to the development of resources and technologies for this sector, aiding design, creating content, testing, providing feedback on pilot versions of applications, ensuring dissemination and take-up of outputs.

Translation Networks: There is an extensive network of translators who provide Irish-language versions of documents from the literary end of the spectrum to the highly technical documents that emanate from State departments and now, increasingly from the EU. These networks have a central role to play in using, testing and evaluating machine translation systems that are under development. The translators in turn stand to benefit in that even partially successful machine translation systems can greatly speed the translation process, improving its efficiency and removing repetitious tasks. The research groups developing machine translation will provide an important hub for grass-roots knowledge-sharing in the use of resources and for training on new platforms.

Disability and Access Groups: the Digital Plan aims to make Irish accessible to many who would otherwise be excluded. Collaboration with disability networks is key to ensuring the technology is adapted to and reaching those who need it. In past developments, such as the Irish screen-reader for the visually impaired [11], such collaborations were paramount (Chapter 15). People with disabilities are often isolated and may be unaware of the potential of the technology and of their rights in this regard. Networking with individuals, teachers, therapists and with specific disability organisations is crucial, to advocate for and to respond to the needs of specific cohorts. This should be a fluid two-way process: users' input is needed for the design and testing of applications; once ready for deployment, these networks are essential to ensure they are widely taken up.

International Research Networks: Collaboration with the international and national research laboratories and centres of excellence in specific areas targeted by the Digital Plan is important to ensure that its work remains at the cutting edge. There are also many European networks that aim to share resources and corpora, and these can be of particular importance in certain areas, for example, in machine translation, where the *ELRC* [12] are improving automated translation for public administration across Europe. It is also important to maintain alignment with the activities of groups such as *CLARIN* [13] and networks such as *META-SHARE* [14] and the European Language Grid. Of particular relevance is *SiGUL* [15], which fosters contact with colleagues working on technologies for underresourced minority and endangered languages.

16.10 Commercial considerations

As the Digital Plan is for the benefit of the Irish language and the language community, and as the research is funded by the State, it is a basic principle that research outputs will be made freely available to the public. As technologies mature, there will be growing opportunities for entrepreneurial ventures to commercialise specific applications, e.g. through the establishment of campus companies. As mentioned in §16.7 above, the possibility of Gaeltacht-campus co-location would be a desirable option. While commercial ventures will not happen immediately, it can be planned for by: including entrepreneurship training in the education of students and researchers; discussing the potential for use of Gaeltacht locations with *Údarás na Gaeltachta;* consideration of policy and guidelines regarding the portion of profits to be fed back to support the research centre involved in the product's development, to further its research on the Digital Plan.

It is hoped that the outputs of the digital plan will percolate beyond the current Irish language community. Small and medium enterprises, whose primary goal is not Irish language-based, will nonetheless see the economic benefit in using the proposed speech and language technologies to integrate Irish into their platforms. There are a number of multinational companies located in Ireland that focus on speech and language in the digital space. As part of their corporate social responsibility, it would seem highly appropriate that such companies would contribute to funding and supporting the Digital Plan's development of a local Irish speech and language technologies sector. In addition to direct funding of the Digital Plan, other forms of support might also be helpful, e.g. localisation of services or websites, embedding Irish proofing tools, web-readers and other applications in their online systems, using Irish (as well as English) in their social media advertising, sponsoring outreach activities that raise awareness of Irish speech and language technologies among the public and the young.

16.11 The Digital Plan as an exemplar for endangered languages

As an endangered minority language, Irish is fortunate in being recognised as the first national language and having State support, as evidenced by the espousal of a Digital Plan for Irish. It has been emphasised throughout that the Plan's development needs to be community-based, firmly rooted in an understanding of the language and of the sociolinguistic context and focused on the particular challenges for Irish in today's world. This is all important to ensure that the core technologies, applications and basic resources are developed in a way that will truly support the language community, and help maintain the language as a living, spoken medium of everyday communication.

The emphasis here on the local dimension does not diminish its universal relevance: we suggest that for all endangered languages, an understanding of the language itself and of the local context is necessary to ensure that the technologies will positively impact on its survival. Although each endangered language is unique in structure and exists in a very specific sociolinguistic context, the kinds of challenges faced tend to be similar, and the solutions adopted in one case can serve as a model for another. The work carried out to date on Irish has greatly benefited from the experience and generosity of researchers working with other minority and/or endangered languages. By implementing the Digital Plan, Ireland can emerge as playing a leading role and ensure that the experience, expertise and, where appropriate, resources are shared with other groups who work to support their endangered, minority and under-resourced languages.

References

[1] Department of Tourism, Culture, Arts, Gaeltacht, Sport & Media. *20-Year Strategy for the Irish Language 2010 – 2030.*

[2] Ní Chiaráin, N. & Ní Chasaide, A. (2015). Evaluating synthetic speech in an Irish CALL application: influences of predisposition and of the holistic environment. In *Proceedings of SLaTE 2015: 6th Workshop on Speech and Language Technologies for Education*.

[3] Government of Ireland. *Official Languages Act*. http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html

[4] Department of Tourism, Culture, Arts, Gaeltacht, Sport & Media. (2018). *The Action Plan for the Irish Language, 2018-2022.*

[5] Government of Ireland. (2005). *Disability Act*. http://www.irishstatutebook.ie/eli/2005/act/14/enacted/en/html

[6] National Folklore Collection, UCD. (2022). *The Schools' Collection*. https://www.duchas.ie

[7] Fiontar & Scoil na Gaeilge, DCU. (2022). *Logainm.ie*. https://www.logainm.ie

[8] O'Reilly, Q. (2014). Want to use Gmail as Gaeilge? Now you can!. *The Journal*. https://www.thejournal.ie/gmail-as-gaeilge-1828979-Dec2014/

[9] An Ríomhacadamh. (2017). [WhatsApp localisation for Irish]. http://www.teicnangael.com

[10] The ABAIR Project. (2022). *MíleGlór*. [Crowdsourcing platform]. https://www.abair.tcd.ie/studio/ga/recorder/

[11] McGuirk, R. (2015). *Exploration of the Use of Irish Language Synthesis with a Screen Reader in the Teaching of Irish to Pupils with Vision Impairment*. [M.Phil. thesis]. Trinity College, Dublin.

[12] European Commission. (2020). *European Language Resource Coordination (ELRC).* https:// www.lr-coordination.eu

[13] *CLARIN - European Research Infrastructure for Language Resources and Technology*. https://www.clarin.eu

[14] *META-SHARE.* http://www.meta-share.elda.org

[15] *SIGUL: Special Interest Group: Under-resourced Languages.* [Joint special-interest group of the European Language Resources Association (ELRA) and of the International Speech Communication Association (ISCA)]. www.elra.info/en/sig/sigul/

Appendix A

In this appendix, we provide a more comprehensive description of the various corpora that have been developed to date for Irish (highlighted in Table 3.1 and Table 3.2) and the immediate requirements identified in terms of future development.

A1. Written Corpora

A1.1. NCI Nua-Chorpas na hÉireann / The New Corpus for Ireland – Gaeilge

Corpus Nua na hÉireann – Gaeilge is a language resource consisting of 30 million words of written texts spanning different styles and sources, consisting of books, newspapers, magazines, textbooks, legislation and manually verified websites [1]. This corpus is an extended version of the earlier Corpas na Gaeilge developed in ITÉ. The corpus has been automatically annotated with part-of-speech (POS) tags (e.g. noun, verb, adverb etc. and other morphosyntactic features such as case, number, gender, tense, person etc.) using a rule-based POS tagger [2] (over 96% accuracy on average) The corpus contains detailed meta-data.

Required Future Developments

- In order to reflect current language usage and to capture the many lexical items that have come into the language since this data was produced (1995-2003), it is essential that new data be added to the corpus.
- A corpus of 30 million words is relatively small by international standards. Foras na Gaeilge report that it needs to be substantially increased to meet the needs of the current corpus-based Irish-English dictionary.
- The spoken element of this corpus is quite small (240K words) [3] and needs to be substantially increased to capture dialectal vocabulary and the many features of discourse not found in written corpora.

A1.2. Gold-Standard Word and Phrase Tagged Corpora

A gold-standard POS-tagged corpus (3,000 sentences randomly selected from the NCI corpus) was created to evaluate the accuracy of the rule-based finite-state part-of-speech tagger for Irish (see Chapter 5). It was used as training data in machine-learning algorithms to learn morphological features and lemmatization (headword) classes [4]. It was also used in developing a rule-based grammatical dependency tagger and chunk parser [5] and it provides the basis for developing the Irish Dependency Treebanks [6, 7] described in the next section.

However, compared to the most-widely used gold-standard POS-tagged corpus for English, e.g. Penn Treebank [8], which has over 4.8 million words with manually corrected part-of-speech tags, the current gold standard POS-tagged corpus for Irish is significantly smaller (115K words), thus highlighting the immediate need for expansion.

Required Future Developments

While current gold standard corpora are rich references for analysing Irish linguistic phenomena, their small size does not lend well to statistical NLP approaches such as language modelling or machine learning – methods that allow a computer to learn patterns of language through observation and probabilistic calculation of linguistic features. For these types of approaches to work, much larger corpora are required. The size of the Penn treebank, for example, has shown this approach to be highly successful. Therefore, the following steps must be taken in relation to Irish POS-tagged corpora:

- Expansion in size of the Irish gold POS-tagged corpus
- Where possible, labelling tasks should be facilitated by semi-automated bootstrapping approaches, followed by manual verification.
- Development of POS-tagged corpora of text from additional domains (e.g. An Vicipéid, web data, social media data, scientific articles, learner corpora, etc.) is also required.

A1.3. Irish Dependency Treebanks

From an NLP perspective, a treebank (see Section 3.4.1) serves as data from which computers can learn linguistic patterns to produce a statistical parser and subsequently process the grammatical structure of previously unseen Irish sentences. A parser (see Chapter 5), is a valuable resource for applications such as statistical machine translation systems, question answering systems, summarisation tools and grammar checkers. However, the current Irish treebanks are relatively small in size.

The Irish Dependency Treebank (IDT) [6] currently contains 1020 sentences (parse trees). This treebank is accompanied by annotation guidelines that clearly describes a dependency analysis of the Irish language and instructions for future annotators to follow. The current size (with only 23K tokens) is significantly smaller than the majority of other standard treebanks (e.g. Arabic 200K+ tokens, Czech 1,500K+ tokens, Spanish 400K+ tokens, Ancient Greek 250K+ tokens).

The Irish Universal Dependencies Treebank (IUDT) is a version of the IDT that has been converted to a "universal" annotation scheme as part of the Universal Dependencies Project[1] [7]. The latest v2.10 release included 4,910 syntactically annotated trees (40K tokens)[2]. This project oversees the development of a variety of language treebanks (currently 220+), used as a basis for cross-lingual parsing and analysis. This project enables low-resourced languages to leverage treebank data and linguistic information from better-resourced languages, as a method of overcoming a lack of data. The inclusion of the IUDT in the UD project has raised the profile for Irish in NLP studies and has also allowed other researchers to use our data sets in cross-lingual parsing or typological studies. The benefits of releasing the Irish data have already become evident, as the language is of much interest in some studies due to its unique linguistic features, which present new challenges to this community [9]. Other work has reported on the inclusion of Irish in cross-lingual analysis for low-resourced languages [10] as well as parsing shared tasks [11].

Designing and building such a resource requires a gold standard POS-tagged corpus, extensive linguistic analysis and annotators who are skilled in the language and understand the computational methods for parsing. The task is time consuming – depending on the complexity of a sentence, annotation of a sentence can take anywhere between 5 minutes to one hour.

Required Future Developments

The current accuracy of the Irish Dependency parser is 71.4% (3) and the Irish UD parser is 71% [12], (84% with gaBERT [35]) (compared with state-of-the-art UD parsers for French 91%, Czech 93% and English 89% [36]). The performance of a parser directly correlates with the size of a treebank. A much larger treebank is, therefore, needed to facilitate future statistical parsing efforts, to ensure the quality of other downstream applications (e.g. MT, information retrieval, grammar checking, CALL, text summarisation).

- The expansion of the Irish Dependency Treebank (IDT) should be concurrent with the expansion of the gold-standard POS-tagged corpus.
- The development of the Universal Dependency Treebank (IUDT) should meet annual releases as per the UD project's schedule ensuring that Irish data continues to be used in studies worldwide.
- Bootstrapping methods should be employed where possible to speed up the development of these resources.
- Crowdsourcing approaches (see Chapter 16 for discussion) should be explored in treebank development (e.g. using gamification methods to combine language learning, gaming and resource building in order to annotate the treebank, as in the case of French annotation)[3].
- The IDT is lacking in comprehensive annotation of multi-word expressions, which are syntactic constructions that play a crucial role in accurately processing language [13, 14]. The development, in its early stages, of a multi-words lexicon for Irish [15] has helped to inform research in this area [34]. Further work is required in this area in alignment with the workings of the UniDive EU COST Action[4].

A1.4. Irish Crúbadán Web Corpus

The Irish Crúbadan Corpus developed as part of the Crúbadán Project [16] is a comprehensive corpus of Irish texts crawled from the web, including news articles, blog posts, social media posts, government and legal documents, literature, etc. After cleaning, removing duplicate sentences, and text written in other languages, the corpus contains over 100 million words of Irish. This corpus is important in part because of its large size, which facilitates the development of language technologies that rely on statistical language modeling.

The Crúbadán corpus drives the development of a number of important language technologies, among them, the GaelSpell spell checker , the grammar checker An Gramadóir , the InterGaelic machine translation engines from Manx and Scottish Gaelic to Irish [17] and the Irish language Caighdeánaitheoir [18]. Subsets of the web corpus have also been used by the lexicographers working on the New English-Irish dictionary project and the Coiste Téarmaíocht as a method for identifying newly coined Irish words.

Required Future Developments

It is important that this crawled corpus continues to be updated daily as new texts appear on the web, and therefore provides an accurate snapshot of the language as it is used by speakers today.

A1.5. An Vicipéid: Irish Wikipedia Online Encyclopaedia

Wikipedia is an online multilingual encyclopaedia developed through the Wikimedia Foundation. The growth of the Irish Wikipedia (An Vicipéid) is an important facet in the provision of language text resources. Not only does Irish Wikipedia provide increased text data for NLP research and development, but also, through the various additional meta-data made available through the wiki-structure, (keywords, hyperlinks, etc.), such a corpus is also an

[1] http://universaldependencies.org
[2] https://github.com/UniversalDependencies/UD_Irish-IDT
[3] http://zombilingo.org/
[4] https://www.cost.eu/actions/CA21167/

invaluable basis for language processing tasks such as named-entity recognition (Chapter 5), sentiment analysis (Chapter 5), natural language generation (Chapter 6), and so on. There is also an educational benefit to language learners in creating of text references on a variety of topics and genres in Irish. Finally, bilingual speakers can benefit from accessing this universally used repository of knowledge through Irish, as opposed to accessing the content through English only.

The Irish Wikipedia (Vicipéid)[1] is currently ranked 93rd in the list of all language wikipedias, with 57,000 articles. To put this into context, Welsh ranks 41st and other more highly ranked languages include Afrikaans (68th), Asturian (64th) and Breton (82nd).

Required Future Developments

- Much work is required to increase the number of Irish language Wikipedia editors and expand the number of articles in An Vicipéid.
- Collaboration and cooperation is required between various stakeholders such as Wikimedia Community Ireland, the Wikimedia Foundation, educational bodies (universities and schools), the NLP community and Irish language experts in coordinating a strategic approach to developing An Vicipéid to ensure a high-quality publicly available resource. Successful models of wikipedia expansion from other language communities such as Welsh[2], Breton, and Basque include providing dedicated secondary school training, and machine translation of content from other language wikipedias (for post-editing by editors).

A1.6. Gaois Corpus of Contemporary Irish (Corpas na Gaeilge Comhaimseartha)

Gaois Corpus of Contemporary Irish (25+ million words) is a collection of Irish-language texts in digital format made accessible for queries or research purposes by Fiontar agus Scoil na Gaeilge (DCU) on Gaois.ie. It consists of edited texts which have been published from the beginning of the 21st century onwards. The corpus currently includes media articles, pieces of literature, and academic writing. All of the corpus has had header data regarding the publisher and genre of text manually added, with approximately 40% of this has been marked up as XML in accordance with TEI (Text-encoding initiative) standards[3]. Work has commenced on the linguistic annotation of the corpus for POS tags, lemmatisation, and morphological information, however, it is at an early stage. Additional texts are periodically collected and added to the corpus. New sources of data are actively sought by the Gaois Research group.

Required Future Developments

- The corpus currently contains approximately 25 million words but needs to be expanded in a balanced manner to ensure it is kept up-to-date with current language usage.
- Linguistic annotation such as POS tags, lemmatisation and morphological information is also required to ensure broad use in language technology and linguistic research.

Where copyright restrictions prevent public release of corpus data, every effort should be made to provide by-products of this resource such as pre-trained word embeddings or language models.

A1.7. Gaois Parallel Corpus

Parallel corpus of aligned text segments from Irish and European Union legislation.

26 million Irish words and 24.5 million English words.

A1.8. Corpora of Irish Social Media Text

A corpus of POS-tagged Irish tweets containing 1,500 random tweets was used as the basis for the TwittIrish Dependency Treebank [19], forming training corpora for Tweet-specific POS taggers and a dependency parser. Such tools are useful for integration into downstream applications that enable analysis of Irish tweets [4, 19]. In addition, both corpora have been expanded with annotations of English code-switching [19, 20]. Researchers at DCU have also collected a small corpus of 1000 parallel English-Irish tweets relating to the Brazil 2014 World Cup [21]. These tweets are the beginnings of a corpus that can be used to build systems to analyse sport-related Irish-language Twitter content.

Required Future Developments

The existing work in this area already provides the basis of sociolinguistic studies on the new variation of Irish language that is evolving online, and provide insights to a new generation of Irish language speakers, and the various ways in which people can learn and engage with the language. This is an area of research that will require much more extensive study.

- The social media corpora will require significant expansion and updating to reflect current use of the language online.
- Further research is required into the use of code-switching in social media between English and Irish online, and as such this will require a larger corpus.
- Future studies on the use of hashtags or

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[2] https://blog.wikimedia.org.uk/2012/07/working-with-the-welsh-language-wikipedia/
[3] https://tei-c.org/

categorisation of Irish language tweets will require the collection of specific types of tweets as opposed to general domain tweets.
- Further research is required to improve the accuracy of the Twitter POS tagging.
- Parsing (extraction of sentence structure) of Tweets requires further research.
- In addition, investigation is required into Irish language use on Facebook and other similar social networks (i.e. not only on Twitter).
- Tools for analysing the sentiment of online content require tailoring to Irish to ensure that the opinions expressed by the Irish language online community are equally taken into account; for which preliminary studies have been carried out [22].

A1.9. Multiword Expression tagged Corpora

Multiword expressions (MWEs) represent a string of two or more words that convey a single meaning (e.g. *buail le* [lit. hit with] 'meet'; *déan iarracht* 'make an attempt'; *cuir fearg (ar)* [lit. put anger (on)] 'make angry'; *bain amach* [lit. extract out] 'earn'). These expressions notoriously cause challenges for natural language processing (Chapter 5) as they are often idiomatic in meaning and difficult to decipher based on the individual words alone. Research has begun in the area of automatic processing of Irish MWEs with the development of a categorisation scheme of MWEs along with the creation of MWE-tagged corpora that form the basis of test and training data for MWE-aware tools [38]: (i) The PARSEME Irish MWE corpus consists of 1700 sentences originally from the Irish UD Treebank [7], annotated with 662 verbal MWEs from seven categories: verbal idioms, inherently adpositional verbs, inherently reflexive verbs, light-verb constructions, and verb-particle constructions. (ii) In order to explore the treatment and effect of MWEs in Irish-English and English-Irish machine translation, an MWE parallel corpus (of 835,867 bilingually aligned sentences) was created and automatically tagged with MWEs for both English and Irish. [iii] A gold test MWE corpus for Irish and English was formed by combining 25 sentences from each of the four domains represented in the data: technical, legal, web and general domain.

A1.10. EduGA: A Corpus of Irish Medium Educational Materials

As a considerable portion of the population of Ireland depends on the education system in order to learn the Irish language [23] the type of language used and taught in this domain is very important to the future of the Irish language. A 7.5 million word corpus of textbooks and educational materials from all levels of education has been compiled and annotated as part of PhD research in Trinity College Dublin. This research focuses on language complexity measures for the domain of education, in order to compare the complexity of language used in materials across different levels of education (e.g. primary level and post-primary level), and also to analyse language use in subjects taught through the medium of Irish [24].

Required Future Developments

- This corpus requires continuous updating in order to reflect the current educational needs and the evolving landscape of language use in educational domains.
- In order for materials creators (including both commercial publishers and teachers) to benefit from this research an API, or web app, needs to be made available. Materials creators could input a lesson they have created, which is then processed in a number of ways, and objective feedback is provided.
- Comparative analyses between learner-generated language and the contents of the materials from which they learn is required. For example, an analysis of the syntax, grammar, and vocabulary in textbooks as opposed to learner essays could be undertaken, in order to provide focussed lessons tailored to the needs of a particular group of learners.
- Further development of the complexity and readability metrics developed in this research in order for Irish to be on a par with other European languages.

A1.11. TEG Learner Corpus of Written and Spoken Irish

A Corpus of Learner Irish is currently being designed, collected and annotated as part of PhD research in the Centre for Language and Communication Studies, Trinity College Dublin, in collaboration with the Centre for Irish Language Research, Teaching and Testing in Maynooth University. The corpus currently consists of 200,000 words of spoken and written learner data, from the Teastas Eorpach na Gaeilge exams, at intermediate to advanced proficiency levels. The corpus is manually transcribed and annotated for linguistic errors and corrections. This followed by automatic annotation with POS tags [25]. Semi-automatic error-detection is also being investigated. This corpus will be analysed to provide clearer, language-specific descriptors of CEFR proficiency levels B1-C1. This analysis will also support the development of empirically-based teaching materials, learning resources and assessment tools tailored to the needs of learners at each proficiency level.

Required Future Developments

- This corpus will provide a methodological foundation for the development and analysis of learner corpus data in Irish, but requires significant expansion in terms of size, learner age and proficiency level, and genre of text in order to facilitate the comprehensive study of learner language in Irish.

- In particular, it is essential that more spoken learner, data, particularly naturalistic interaction (e.g. classroom and teacher language) also be included in the Learner Corpus.

- Further work on the automation of error detection and annotation is necessary to facilitate the expansion of this learner corpus, as well as the development of targeted error correction tools for teachers and learners of Irish.

A1.12. National machine translation corpus and the ELRC Parallel Corpus

Relative to other EU official languages, publicly available parallel data sets for Irish are limited. In 2015, first steps were made towards a national collection of such data in Dublin City University's development of the Tapadóir SMT system for Ireland's Department of Tourism, Culture, Arts, Gaeltacht, Sports and Media (DTAGSM) [26]. This includes a web-crawled corpus [16], Irish legal documents, newly crawled web data and translation memory files from DTAGSM. Since then the system has been updated with data from other sources such as Údarás na Gaeltachta, the Office of the Irish Language Commissioner, city councils, universities, and some government departments. Monolingual data from Foras na Gaeilge has also proved valuable for language modelling in this system development. This data collection will also contribute to the development of any translation systems for use within the proposed national Shared Translation Service (see Chapter 11).

The Tapadóir R&D project has been complemented by activities of the European Language Resource Coordination (ELRC) project[1], which is a European Commission initiative for collecting translation data across all European languages in order to improve the EU's eTranslation system. As a member of the ELRC, Ireland has two National Anchor Points; a public administration representative in DTAGSM and a technical representative at the ADAPT Centre, Dublin City University. Their involvement has prompted a nationwide collection of public administration translation data for the purposes of improving translation tools at the European and national level. Through this, a report was produced on the management of translation data and use of translation technology in Irish public administration [37].

In 2019, under EU-funding, the National Relay Station (online portal) was developed to facilitate this data collection. The portal was relaunched in 2022 as eSTÓR[2]. Access to this data portal is granted to all those working with Irish language data in public administration. The site facilitates the generation of translation memory files from batches of uploaded parallel translated documents. In addition, through group sharing, users are able to access language data (e.g. translation memory files) from other users that will help their own translation efforts, and in many cases reduce future translation costs. The majority of the translation data uploaded to the portal by those in public administration is freely available for download under the EU Open Data Directive, supported by the national Open Data Unit.

A more recent EU-funded project at DCU, PRINCIPLE, focused on collecting data for languages that were identified by the European Commission as "low-resourced" in terms of data required for building machine translation engines (Irish, Croatian, Icelandic and Norwegian). Through this project, Irish stakeholders benefitted from assistance with collection, processing, cleaning and alignment of existing or archived translation data in order to make it useful for translation re-use and the development of translation technology. Contributors and benefactors in this project included Rannóg an Aistriúcháin, Foras na Gaeilge, NUI Galway and the Department of Justice, and covered the domains of eJustice and eProcurement. This project served as a prime example of how sharing well-managed and curated translation data can enhance translation activities through both time and cost reduction within a public organisation.

Required Future Developments

- Ongoing administrative support and technical maintenance of eSTÓR is required to ensure that translation data collection and management continues at a national level.
- Continued cooperation with the European Commission is required in the ELRC-led drive for data collection (under the Connecting Europe Facility (CEF)[3] Automated Translation programme and the EC's Common European Data Spaces[4]) aimed at improving automated translation in public administration across all EU countries (e.g. Irish Government Shared Translation Services).
- Improving web-data crawling for English-Irish data is also an immediate need. Closer collaboration with web-crawler developers (e.g. ILSP web-crawler [16, 27] is required to improve methods for scraping English-Irish text from .ie domains. However, the Irish crawled corpus of the Paracrawl project [28] (19) contains machine-translated content along with much data already collected through ELRC methods.
- There is also a need for exploring methods

[1] http://www.lr-coordination.eu
[2] https://estor.ie
[3] https://ec.europa.eu/digital-single-market/en/connecting-europe-facility
[4] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

in extracting relevant aligned phrases from comparable corpora for inclusion in parallel text data sets. This could be deemed an effort to overcoming the lack of Irish corpora resources while maximising the text that is actually available.

- Closer cooperation is required with the Department of Public Expenditure and Reform's Open Data Unit and the Office of Government Procurement in creating awareness of the value of translation data and the benefit to the state through improving data-sharing activities.

A1.13. National Folklore Collection – Schools' Collection

As part of the Meitheal Dúchas.ie project to digitise the National Folklore Collection, the Schools' Collection manuscripts are being transcribed by members of the public via the web. The results of this work now comprises a crowdsourced corpus of over 3.8 million words in Irish from the 1930s which can be searched via Dúchas.ie. The website received 450K unique visitors in 2019.

Required Future Developments
The search facility in the Dúchas website requires much improvement. Recent preliminary study showed the possibility of including language technology to facilitate searching the database [29].

A1.14.Corpas Stairiúil na Gaeilge 1600-2000

Corpas Stairiúil na Gaeilge is a comprehensive corpus of the Irish language stretching back 400 years. The corpus is being developed in the Royal Irish Academy as the basis of Foclóir Stairiúil na Gaeilge, a research project whose aim is the production of an historical dictionary of Irish language for the period 1600-2000.

The corpus consists of five major sub-corpora representing different historical periods of Modern Irish. To date sub-corpora for the periods 1600-1882 and 1882-1926 have been compiled and annotated and are available online[1]. As this is a historical corpus, data capture for the earlier time periods require the use of optical character recognition (OCR) and manual checking, together with (manual and automatic) language standardisation and automatic POS tagging [30].

Required Future Developments
Work on subsequent periods of the corpus is required to enable completion of Corpas Stairiúil na Gaeilge and to facilitate work on the dictionary itself.

A1.15.Corpas Filíocht Shiollach na Gaeilge 1200-1650

This corpus consists of almost two thousand poems written in Classical Irish during the period 1200 to 1650. Currently five hundred of these poems and extensive metadata have been made available online to scholars and the public. See https://www.tcd.ie/slscs/research/areas/corpora/bardic.php for further details.

Required Future Developments
Further research is required to improve lemmatisation and POS tagging for texts of this period, in order to make the remainder of this corpus publicly available.

A2. Speech and Spoken Corpora

In this section, we summarise a number of speech corpora (scripted speech) and spoken corpora (spontaneous speech).

A2.1. ABAIR General Speech Synthesis Corpus

Corpora are continuously being designed and recorded for the synthetic voices of the different Irish dialects within ABAIR. The general principle is to use dialect-appropriate materials for recording and to strive for good phonetic coverage of sounds in all contexts. The size of these corpora is highly variable, reflecting the requirements of the specific speech engine used to build a voice (see Ch. 8). To date, the focus has been on the 3 main dialects but work on further sub dialects is underway. There is currently c. 25 hours of recorded speech: selected portions of these corpora have been optimised for synthesis – edited, annotated for corrected for hesitations etc., segmented, aligned to the phonetic transcription (X-SAMPA, IPA) with stress marking etc.

Required future developments
As described in Ch 8, extensive further corpus design, recording and preparation/annotation will be required, to cater for

- Full coverage of the Irish dialects, including the most endangered. This includes collecting dialect-specific materials and ongoing analysis of the phonetic coverage of corpora
- More speakers, and especially children's voices
- Corpora for eventual code-switching and bilingual synthesis
- Prosodically rich corpora to provide
  - for a wider range of synthesis applications (less monotonous voice)
  - expressive, interactive speech corpora (containing emotion, mood, attitude)

[1] http://corpas.ria.ie/

These corpora should improve the quality of all synthetic speech output but are critically important for conversational dialogue systems

A2.2. ABAIR Compact synthesis corpus: An Corpas Beag

An Corpas Beag, refers to a suite of corpora which are under development in the TCD ABAIR group. These are compact corpora for the individual dialects of Irish which aims to provide maximal phonetic coverage with the least amount of scripted prompting material. Such a corpus has been designed for the Munster dialect (Corpas Beag na Mumhan) and for the Connaught dialect (Corpas Beag Chonnachta) and work is ongoing on similar compact corpora for the other dialects of Irish.

Required future developments

Requirements are as described for the general synthesis corpora of §2.1 above, with the additional need to:

- Create content that is dialect appropriate and provides full phonetic coverage.
- Record, annotate and refine corpora for optimisation in specific speech engines

A2.3. Comhrá Spoken Corpus

The Comhrá Corpus [1] currently consists of 240K words of transcribed spontaneous speech, and includes over 200 speakers from all of the major dialects of Irish. This spoken corpus was designed to give balanced coverage of the dialects and other sociolinguistic features such as age and gender. The current corpus represents a subset of the original corpus design[2].

Currently the corpus contains over one hundred transcribed audio files each of eight minutes duration from a variety of sources including Raidió na Gaeltachta (RnaG) podcasts, RTÉ archival audio files, and some samples from the Hartmann Spoken Corpus of Connemara Irish which was recorded in Ros Muc, Galway in the 1960's [31]. In addition, there are a some recordings of spontaneous interactions, which were recorded in four Gaeltacht locations and in Dublin. The Comhrá corpus has been POS-tagged with a rule-based finite-state POS tagger which has been enhanced to handle features which are specific to spoken language [3]. The tagger incorporates an English language module in order to accurately tag English lexical items and code switching. The current corpus, funded by Foras na Gaeilge for the New English-Irish dictionary project[3] is quite small (240 K words) compared to spoken corpora for other languages such as English, e.g. the British National Corpus[4] contains 10 million words of transcribed spoken data, which represents 10% of the 100-million word corpus, and the International Corpora of English[5], all of which contain 600K words of spoken English.

Required Future Developments

- The Comhrá spoken corpus is incomplete in many of the areas specified in the initial corpus design. The corpus needs to be significantly extended, particularly through the addition of specific domains of language use, e.g. business settings, classroom discourse, hair salons, supermarkets etc.
- In the case of spoken corpora, particular attention needs to be paid to the quantity of data to be recorded and transcribed, the number of subjects to be recorded, their linguistic background, age, dialect, as well as the depth to which the corpus is annotated. Additional factors to be considered include the quality and type of recording, the precise recording equipment and recording environment, the nature of the materials to be recorded.
- Existing transcribed data (e.g. RnaG podcasts) can incorporated by aligning the transcript with the audio file if available.
- A spoken corpus is a valuable resource for language teaching and learning. For maximum benefit, the audio transcripts should be categorised according to various pedagogical categories, such as subject matter, speed and complexity of the speech, clarity of diction, etc..

A2.4. An Scéalaí - An Corpas Cliste

This is a learner corpus being collected as part of an iCALL platform which is currently under development. The goal of An Scéalaí is to encourage written production from learners, while exposing them to spoken renditions of their own text (using the synthetic voices of ABAIR) and eliciting spoken output. As part of the platform, learners provide information on their sociolinguistic background, along with a self assessment of their language level, which in turn is matched to their written and spoken output. All data is anonymised

Required future developments

These are very novel learner corpora and as they grow, tools and frameworks will need to be developed so that they can be analysed to inform second language acquisition research and the priorities for future CALL development, including future iterations of An Scéalaí.

[1] https://www.scss.tcd.ie/~uidhonne/comhra/index.utf8.html
[2] https://www.tcd.ie/slscs/research/projects/past/gala.php
[3] https://www.focloir.ie/
[4] http://www.natcorp.ox.ac.uk/corpus/index.xml
[5] https://www.ucl.ac.uk/english-usage/projects/ice.htm

A2.5. ABAIR MíleGlór Speech Recognition Corpus

The MíleGlór corpus is currently being collected for the development of Irish speech recognition systems. A targeted crowdsourcing approach is being used where the public are contributing via phone and computer at https://www.abair.tcd.ie/mileglor. Field recordings are also taking place in Gaeltacht locations, with small samples of speech being collected from large numbers of speakers, covering all dialects, all ages, all levels. This work is being coordinated with the corpus collection for speech synthesis (See Chapters 8 and 9). Relevant sociolinguistic information is noted prior to a recording session and this information is aligned to an identification code associated with the recording.

The MíleGlór corpus is being used alongside numerous pre-existing speech data, accompanied by text (e.g., from media sources).

Required Future Developments

- Extensive recordings, extending the current coverage.
- Collection of other existing materials that can be used
- Corpus cleaning (prooflistening to e.g., remove extraneous material, label hesitations etc.)
- A stable platform to manage the online recordings from any source. This could include a specifically tailored iOS and android app for recording, which would be easy to use
- Server and management to ensure the data is hosted, protected and effectively used in the building of speech recognition systems (see Ch 9), and applications.

A2.6. TEG Learner Corpus of Written and Spoken Irish

The corpus currently consists of 200,000 words of spoken and written learner data, from the Teastas Eorpach na Gaeilge exams, at intermediate to advanced proficiency levels. See section 1.10 TEG Learner Corpus of Written and Spoken Irish for further details.

A2.7. International Comparable Corpus: Spoken and Written Irish

The aim of the International Comparable Corpus (ICC)[1] project is to create a set of comparable corpora in a number of languages including Irish (other languages include Czech, English, Finnish, French, German, Italian, Norwegian, Polish, Slovak, and Swedish). Each corpus will consist of 1 million words, comprising of 40% written language and 60% spoken language. The ICC corpora uses the ICE corpus design for maximum comparability with existing corpora for English (including Irish English) and will re-use existing language resources wherever possible.

An important and unique feature of each ICC corpus is its substantial spontaneous spoken component, comprising ca. 600,000 words (60% of the corpus). Such provision of spoken data across 15 or more discourse situations for contrastive analysis among several languages will allow the much-needed and unprecedented cross-linguistic corpus-based comparisons of spoken language [32]. Along with balanced data across the written registers, ICC will be an invaluable resource for machine-learning, e.g. MT [33], for corpus-based lexicography and for linguistic research.

Required Future Developments

- Assess the availability of suitable existing written and spoken resources for use in ICC. It is likely that there will be sufficient written resources for the written component (400 K words). However, it is expected that the majority of the 600K spoken part of the corpus will need to be recorded, transcribed and annotated. While there are ample broadcast audio files (podcasts) and transcriptions available, e.g. in the Comhrá spoken corpus (see section 2.3 etc.), this domain of use represents only 10% of the spoken corpus design. In contrast, there is very little spontaneous spoken data for the other discourse situations which account for 90% of the spoken part of ICC.

A3. References

[1] Kilgarriff, A., Rundell, M. & Uí Dhonnchadha, E. (2007). Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation 40*(2), 127-52.

[2] Uí Dhonnchadha, E. & van Genabith, J. (2006). Scaling an Irish FST morphology engine for use on unrestricted text. In A. Yli-Jyrä, L. Karttunen & J. Karkhumäki (Eds), *Finite-State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, 2005. Revised Papers*. Springer, pp. 247-258.

[3] Uí Dhonnchadha, E., Frenda, A. & Vaughan, B. (2012). Issues in Designing a Corpus of Spoken Irish. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 23-25).

[4] Lynn, T., Scannell, K. & Maguire, E. (2015). Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the ACL 2015 Workshop on Noise User-generated Text* (pp. 1-8).

[1] https://korpus.cz/icc

[5] Uí Dhonnchadha, E. (2009). *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. [Ph.D thesis]. Dublin City University.

[6] Lynn, T. (2016). *Irish Dependency Treebanking and Parsing*. [Ph.D. thesis]. Dublin City University and Macquarie University.

[7] Lynn, T. & Foster, J. (2016). Universal Dependencies for Irish. In *Proceedings of the 2nd Celtic Language Technology Workshop* (pp. 79-92).

[8] Marcus, M., Santorini, B. & Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. In S. Armstrong (Ed.) *Using Large Corpora*. Cambridge, MA: The MIT Press, pp. 273-290.

[9] Futrell, R., Mahowald, K. & Gibson, E. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the 3rd International Conference on Dependency Linguistics* (pp. 91-100).

[10] Duong, L., Cohn, T., Bird, S. & Cook, P. (2015). A Neural Network Model for Low-resource Universal Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 339-348).

[11] Zeman, D., Hajic, J. Popel, M. Potthast, M., STraka, M., Ginter, F., Nivre, J. & Petrov, S. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 1-21).

[12] Straka, M. & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88-99).

[13] Baldwin, T. & Kim, S.N. (2010). Multiword expressions. In N. Indurkhya & F.J. Damerau (Eds), *Handbook of Natural Language Processing, 2nd Edition*. Boca Raton: CRC Press, pp. 1-39.

[14] Monti, J. & Amalia, T. (2015). Multiword Units Translation Evaluation in Machine Translation: Another Pain in the Neck? In *Proceedings of Multi-word Units in Machine Translation and Translation Technologies 2015* (pp. 25-30).

[15] Walsh, A., Lynn, T. & Foster, J. (2019). Ilfhocail: A Lexicon of Irish MWEs. *In Proceedings of the Joint Workshop on Multiword Expressions and WordNet 2019* (pp. 162-168).

[16] Scannell, K. (2020). [Software]. http://crubadan.org/

[17] Scannell, K. (2014). Statisical models for text normalization and machine translation. *In Proceedings of the 1st Celtic Language Technology Workshop, 25th International Conference on Computational Linguistics* (pp. 33-40).

[18] Scannell, K. (2009). Standardization of corpus texts for the New English-Irish Dictionary. [Paper presentation]. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[19] Cassidy, L., Lynn, T., Barry, J. & Foster, J. (2022). TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computation Linguistics, Volume 1 Long Papers* (pp. 6869-6884).

[20] Lynn, T. & Scannell, K. (2019). Code-switching in Irish tweets: A preliminary analysis. In *Proceedings of the Celtic Language Technology Workshop* (pp. 32-40).

[21] Centre for Global Intelligent Content, Dublin City University and Trinity College, Dublin. *Brazilator: Machine Translation & Sentiment Analysis for World Cup 2014*.

[22] Afli, H., McGuire, S. & Way, A. (2017). Sentiment translation for low resourced languages: Experiments on Irish general election tweets. [Paper presentation]. *18th International Conference on Computational Linguistics and Intelligent Text Processing.*

[23] Central Statistics Office. *Census 2016*. Dublin: CSO.

[24] Ó Meachair, M.J. (2019). *The Creation and Complexity Analysis of a Corpus of Educational Materials in Irish (EduGA)*. [Ph.D. thesis]. Trinity College, Dublin.

[25] Ní Ghloinn, A., Uí Dhonnchadha, E. & O'Keeffe, A. (2018). The design and annotation of the TEG learner corpus of Irish. [Paper presentation]. *Inter-Varietal Applied Corpus Studies International Bienneial Conference.*

[26] Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivavasta, A., & Judge, J. (2015). Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish. *The 4th LRL Workshop: "Language Technologies in Support of Less-Resourced Languages.*

[27] Papavassiliou, V., Prokopidis, P. & Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora* (pp. 43-51).

[28] Esplà-Gormis, M., Forcada, M.L., Ramírez-Sánchez, G. & Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII, Volume*

2 (pp. 118-119).

[29] Ó Raghallaigh, B., Scannell, K. & Dowling, M. Improving full-text search results on dúchas.ie using language technology. In *Proceedings of the Celtic Language Technology Workshop* (pp. 63-69).

[30] Uí Dhonnchadha, E., Scannell, K., Ní Mhearraí, E., Ní Mhaoláin, M., Ó Raghallaigh, B., Toner, B., Mac Mathúna, S., D'Auria, D., Ní Ghallchobhair, E. & O'Leary, N. (2014). Corpas na Gaeilge 1882-1926: Integrating Historical and Modern Irish Text. In *LREC 2014 Workshop LRT4HDA: Language Resources and Technologies for Processing and Linking Historical Documents and Archives* (p. 12).

[31] Wigger, A. (2000). *Caint Chonamara: Bailiúchán Hans Hartmann*. University of Bonn.

[32] Cermáková, A., Jantunen, J., Jauhiainen, T., Kirk, T., Kren, M., Kupietz, M and Uí Dhonnchadha, E. (2021). International Comparable Corpus: Challenges in building multilingual spoken and written comparable corpora. *Research in Corpus Linguistics* 9(1), 89-103.

[33] Sharroff, S., Rapp, R., Zweigenbaum, P. & Fung, P. (Eds). (2013). *Building and Using Comparable Corpora*. Springer.

[34] MGuinness, S.L., Phelan, J., Walsh, A. & Lynn, T. (2020). Annotating MWEs in the Irish UD Treebank. In *Proceedings of the 4th Workshop on Universal Dependencies* (pp. 126-139).

[35] Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Ó Meachair, M.J. & Foster, J. (2022). gaBERT – an Irish Language Model. *In Proceedings of the 13th Conference on Language Resources and Evaluation* (pp. 4774-4788).

[36] Kondratyuk, D. & Straka, M. (2019). 75 languages, 1 model: parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 2779-2795).

[37] Berzins, A., Choukri, K., Giagkou, M., Lösch, A., Mazo, H., Piperidis, S., Rigault, M., Schnur, E., Small, L., van Genabith, J. & Vasiljevs, A. (2019). Sustainable Language Data Sharing to Support Language Equalty in Multilingual Europe – Why Language Data Matters: ELRC White Paper.

[38] Walsh, A. (2023). *The Automatic Processing of Multiword Expressions in Irish*. [Ph.D. thesis]. Dublin City University.