The Relo-KT Process for Cross-Disciplinary Knowledge Transfer

Transferring linguistic understanding of rhetorical figures to the machine translation domain

A thesis submitted to the
University of Dublin, Trinity College
In fulfilment of the requirements of the degree of
Doctor of Philosophy

Emma Louise Clarke
School of Computer Science & Statistics
Trinity College Dublin, Ireland
clarkee8@tcd.ie

Supervised by Professor Owen Conlan

Declaration of Authorship

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

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed:

_____

Date:

_____

Abstract

Digital humanities research, by its nature, is collaborative and interdisciplinary. A key aim when undertaking cross-disciplinary research is to integrate insights from two or more distinct disciplines using both formal and informal methodologies to achieve effective knowledge transfer. Such collaboration can lead to new ideas, creative solutions and innovation within both disciplines, in a way not possible within single discipline research and work. Whilst collaboration happens regularly in academic, creative and work environments, methods of cross-disciplinary knowledge transfer in interdisciplinary research are not often documented.

This thesis explores synergies between the linguistics discipline and the extensive science around machine language translation. While both disciplines have their own distinct approach to solving problems, combining these disparate skills within a particular application affords exciting opportunities to develop.

The multi-step relo-KT process was developed during this thesis to formalise and codify collaborative cross-disciplinary knowledge exchange. The process incorporates establishing an interdisciplinary question; acquiring a corpus of data suitable for analysis and extracting domain specific understanding from it. The process is iterative in nature as the cross-disciplinary knowledge codification and transfer develops between the discipline experts.

To rigorously examine the relo-KT process, it is applied to the RF-MT (rhetorical figure-machine translation) use case, in which linguistic understanding of rhetorical figures is codified to facilitate a tangible transfer of linguistic knowledge to the machine translation (MT) domain.

A multi-faceted, mixed method approach is taken to enact the relo-KT process. The *Rhetorica* software is deployed to automatically detect rhetorical figures from a corpus of political statements. Key rhetorical figures explored

include epanaphora, epistrophe, polyptoton and polysyndeton, each well-understood within the linguistics field, but each posing challenges for effective machine translation.

Quantitative findings from the application demonstrate the complex nature of persuasive speech. A repository of exemplary codified rhetorical figures for persuasion is developed and improved over a series of semi-structured, collaborative interviews with experts from the field of machine translation. Qualitative findings from the iterative series of interviews indicate that the MT domain is primed to integrate linguistic nuance, and a potential application is in the area of automated post-editing of machine translations.

Acknowledgements

Firstly, I would like to express sincere gratitude to Owen Conlan. I could not have asked for a better supervisor throughout this process. Thank you for being endlessly supportive, positive and generous with your time and advice.

Thanks also to everyone in the ADAPT Centre, and the School of Computer Science & Statistics in Trinity College Dublin, especially Alex O'Connor, who has since moved to pastures new, but was instrumental in shaping my work as a co-supervisor in the early days of my PhD. Special thanks also goes to the ADAPT Centre researchers who agreed to participate in the interviews I carried out in the course of this research.

I am extremely grateful to my examiners, Steven Krauwer and Jennifer Edmond for taking the time to examine my work thoroughly and with insight.

To Susan Schreibman and everyone I worked with while at Maynooth University: it was very exciting to be involved with a project like Letters of 1916. The weekly exposure to the DH world was invaluable to me during my PhD.

I'd also like to acknowledge the support of good friends I've made during my time in TCD especially Vicky Garnett; Angela Griffith and Antonia Hart.

On my first day of the M.Phil in DH & Culture in 2012, Karolina Badz replied to one of my tweets about a module I was taking & that was the start of a really great friendship! Thank you Kar for being a source of constant inspiration.

Special thanks has to go to John, James, Marian and Niall, my team of eagle-eyed proofreaders! Thanks also to my father Seán: without your keen eye,

constructive criticism and general wisdom, the final two weeks of writing would have been way more challenging than they were.

There have been so many supportive ears and shoulders to cry on along the way. I'm so grateful for all of them including Marian, Mary, Elma, Tim, Patrick, Eileen, Claire, Joan, Denis, Brenda, Cathy, Lizzie, Inna, Niamh, Deirdre, The Byrne clan & all the members of the "Fake Book Club".

Thanks to my wonderful family for their constant and unconditional support. Breda, Seán, John, James, Julie, Simon & Saoirse Mae - your kind words, delicious food, hospitality, funny videos, impromptu meetups and general encouragement always arrive at the right time!

Kali and Jarvis deserve a special mention too - for all their meows, cuddles and their fluffy company on the otherwise solitary write-up days!

Finally, none of this would have been possible without my husband & best friend Niall. Thanks for all the fun, adventures and distraction when I needed it most! Your love, support, partnership and good humour kept me going all through this PhD journey.

Contents

List of Figures

List of Tables

Glossary of Terms

Corpus
A collection of texts used for linguistic analysis

Diachronic Corpus
A corpus containing texts from different time periods which is used to study language development or change.

Discourse Analysis
The study of language beyond the sentence level - the study of real language use by real speakers in real situations.

Epanaphora
Repetition of the same word or group of words at the beginning of successive clauses, sentences, or lines (also known as anaphora).

Epistrophe
Ending a series of lines, phrases, clauses, or sentences with the same word or words.

Explicit Knowledge
Knowledge which is easy to articulate and codify.

eXtensible Markup Language
See XML.

Figure of Speech
See Rhetorical Figure.

Interdisciplinarity
Communication and collaboration across academic disciplines.

Iterative Process
A process that is repeated, using the results from the previous stage.

Knowledge Transfer
Sharing or disseminating knowledge and providing inputs to problem solving.

Machine Translation
: The translation of text carried out by a computer, with no human involvement.

Markup Language
: A system for annotating a document for computer processing. See also XML.

Neural Machine Translation
: Neural Machine Translation uses an artificial neural network (a type of framework of machine learning algorithms) to predict the likelihood of the sequence of words in a translation.

Oxymoron
: A figure of speech in which apparently contradictory terms appear in conjunction.

Polyptoton
: A figure of speech in which a word is repeated in close proximity, but in a different form.

Polysyndeton
: A figure of speech which employs many conjunctions between clauses, often slowing the tempo or rhythm.

Post-editing
: The manual correction of texts which have been translated from a source language into a target language by a machine translation system.

Rhetoric
: The art of effective or persuasive speaking or writing, especially the exploitation of figures of speech.

Rhetorical Figure
: A linguistic technique for making a single phrase striking and memorable.

Scheme
: An artful deviation from the ordinary arrangement of words. One of two general categories for figures of speech, along with trope.

Statistical Machine Translation

Statistical Machine Translation involves learning by examining a bilingual corpus and producing a translation based on what it has learned.

Tacit Knowledge

Knowledge which is difficult to transfer to another person by means of writing it down or verbalizing it.

Transcreation

The process of adapting a message from one language to another, while maintaining its intent, style, tone, and context.

Tricolon

A rhetorical figure which consists of three parallel clauses, phrases, or words, which happen to come in quick succession without any interruption.

Trope

An artful deviation from the ordinary or principal signification of a word. One of two general categories for figures of speech, along with scheme.

XML

A metalanguage which allows users to define their own customized markup languages, especially in order to display documents on the Internet.

XML Schema

A description the structure of an XML document (also referred to as XML Schema Definition (XSD)).

Abbreviations

AI Artificial Intelligence.

ALPAC Automatic Language Processing Advisory Committee.

CHI Conference on Human Factors in Computing Systems.

CLARIAH Common Lab Infrastructure for the Arts and the Humanities.

DARIAH Digital Research Infrastructure for the Arts and Humanities.

DH Digital Humanities.

DIKW data-information-knowledge-wisdom.

DL Deep Learning.

EADH European Association for Digital Humanities.

EBMT Example Based Machine Translation.

GIS Geographic Information System.

GNMT Google Neural Machine Translation.

HT Human Translation.

IDR Interdisciplinary Research.

KM Knowledge Management.

KT Knowledge Transfer.

ML Machine Learning.

MLA Modern Language Association.

MT Machine Translation.

NLG Natural Language Generation.

NLP Natural Language Processing.

NLTK Natural Language Toolkit.

NMT Neural Machine Translation.

OECD Organisation for Economic Co-operation and Development.

PMI Project Management Institute.

POS Part of Speech.

RBMT Rule Based Machine Translation.

RF-MT Use case to transfer understanding of Rhetorical Figures (RF) to the Machine Translation (MT) domain.

RO Research Objective.

RQ Research Question.

SMT Statistical Machine Translation.

TD Teachta Dála.

TEI Text Encoding Initiative.

UCD University College Dublin.

UCL University College London.

US United States.

USSR Union of Soviet Socialist Republics.

WAS Waterloo Annotation Scheme.

XML eXtensible Markup Language.

For Saoirse Mae

Chapter 1

Introduction

The transfer of interdisciplinary tacit and explicit knowledge has posed significant challenges for business and academic communities across the globe, especially since the emergence of the concept of the 'knowledge economy' in the 1960s. One of the pursuits of digital humanities scholars since the inception of the field has been to explore the intricate patterns of human language using digital technologies. Linguistic insight has also contributed to the enhancement of automated machine translations with varying degrees of success. Machine translation is the translation of text carried out by a computer, with no human involvement, though human post-editing is typically needed in most machine translation applications. Linguistic subtleties and nuance remain difficult to transfer.

For millennia, rhetorical figures have been utilised, both consciously and unconsciously, for persuasive purposes by many orators and writers. Over 400 such figures have been documented, although a much smaller number are in common usage. Digital tools are now available to identify these rhetorical figures in linguistic corpora.

In a world which is becoming inherently more interdisciplinary (Viseu, 2015), research which involves a synergy of two or more academic, scientific or artistic disciplines is a crucial component of 21st century scholarship (Van Noorden, 2015). Bibliometric analysis of published papers shows that the number of papers which mention 'interdisciplinarity' in the title has been consistently increasing, particularly in social sciences and humanities papers (Larivière and Gingras, 2014). McCarty and Deegan (2016) observe that humanist research practices are changing. Along with a move away from the

traditional solo work associated with humanities researchers, there is an increased trend towards interdisciplinary collaboration. Indeed, one of the central tenets of digital humanities (henceforth referred to as DH) scholarship is its collaborative nature (Edmond, Bagalkot, and O'Connor, 2016; McCarty and Deegan, 2016).

DH is not an easy term to define. The 'What Is Digital Humanities?'[1] website is a collection of 791 quotes attempting to answer that very question from participants in the Day of DH[2] initiative between 2009-2014. Among of the most frequently used words from those definitions are 'new', 'methods', 'tools', 'technology', 'computing' and the terms 'digital' and 'humanities' themselves[3]. What may come as a surprise is the terms 'interdisciplinary' 'collaborate' and 'communicate' are used relatively infrequently[4] when defined by those within the field. A typical definition of DH is the one proffered by Kathleen Fitzpatrick initially in 2010, and again in 2012, that DH could be understood as "a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities" (Fitzpatrick, 2012). It is this nexus, the meeting point of disciplines, that this work focuses on. This thesis considers the interactions which take place when experts from different domains (specified spheres of activity or knowledge[5]) collaborate and communicate to share their expertise.

The concept of **knowledge transfer** involves sharing knowledge, skills and expertise in a formalised way. It is a key part of knowledge management and is present in most sectors in the form of documentation, manuals, guidelines and handovers. These systems serve to take tacit knowledge, which by its nature is difficult to transfer, and codify it in a way which makes it easier to transfer. Knowledge sharing is becoming increasingly important in

[1] What Is Digital Humanities? https://whatisdigitalhumanities.com/

[2] Day of DH is a project that examines the state of the digital humanities through the lens of those within it: https://twitter.com/DayofDH

[3] Day of DH data (2009-2014) was downloaded from: https://github.com/hepplerj/whatisdigitalhumanities. Frequencies were calculated using the open-source, web-based Voyant Tools application: https://voyant-tools.org/. The corpus contained 29,581 words in total. The frequencies for the terms are: humanities (1033), digital (947), tools (230), new (237), technology (182), methods (136), computing (128)

[4] interdisciplinary (32), collaborative (24), collaboration (17), collaborate (6), communication (29), communicate (11)

[5] Oxford Dictionary: https://en.oxforddictionaries.com/definition/domain

moving away from traditional disciplinary silos towards more holistic, cross-disciplinary approaches ("Disciplinary Dilemma: Working across Research Silos Is Harder than It Looks | Andy Stirling").

There is no "one size fits all" definition of interdisciplinarity (Barković, 2010; Holbrook, 2013). Similarly, there is not a particular set of challenges which arises when undertaking interdisciplinary research. Among the challenges which can arise however, is the question of how to communicate the complexity and nuance of subject area expertise from one discipline to another. As technology advances rapidly, one of the emerging issues that faces researchers is how to integrate domain-specific knowledge and understanding in their machine learning systems in a meaningful way. This thesis presents a method which can facilitate such interdisciplinary communication. To establish the validity of the method, it is enacted in the relo-KT process which takes linguistic understanding of rhetorical figures for persuasion and effectively communicates this understanding in a grounded way which can be utilised in a machine translation (MT) workflow.

The remainder of this chapter is comprised of background introductory information about rhetorical figures and machine translation (MT); the motivation for carrying out this research; the research questions and aims and objectives.

1.1 Background

I have a dream today.

Dr Martin Luther King Jr
28 August 1963

1.1.1 Rhetorical figures

In Washington DC, on a sweltering August day in 1963, the civil rights activist Martin Luther King Jr. addressed crowds of approximately 250,000, at the March on Washington for Jobs and Freedom. In what has come to be known as the *"I have a dream"* speech, King advocated for an end to inherent racism in the United States of America. Throughout his address, he repeated the phrase *"I have a dream"*, but he was not simply uttering a slogan. He used the repetition to persuade the demonstrators at the march that the time for change in race relations in the U.S.A. had come. During the speech, King uttered the phrase *"I have a dream"* eight times. The tautology galvanised the large gathering, and it became one of the defining moments of the U.S. Civil Rights Movement. The success of this speech can in part be attributed to King's calculated use of the art of persuasion, in the form of rhetorical figures, to connect emotionally with demonstrators.

The nuances of human speech such as sarcasm, slang and wordplay and our ability to process and understand the subtleties of each make them equally fascinating and frustrating for language researchers. The key focus of this work will be on the subtle linguistic tricks used in persuasive speech, and how these figures can be incorporated into a machine translation process. Nuance is one of the aspects of human speech that is typically missed in the translation workflow. Depending on the context, a loss of linguistic subtlety may have little impact on a translation. However, when translating persuasive texts like King's *"I have a dream"* speech, one of the core elements of the speech is the impact of the repetition of the key phrase on the listener. This repetition often manifests itself in the form of rhetorical figures.

A rhetorical figure is a linguistic technique "for making a single phrase striking and memorable" (Forsyth, 2014). Rhetorical figures are ubiquitous. Where you find communication and persuasion in speech, rhetorical figures can be found concealed in the blend of words. One of the most commonly used figures is *epanaphora*, which is the repetition of a word or phrase at the beginning of successive clauses, famously used by Winston Churchill in his *We Shall Fight on the Beaches* speech delivered in the House of Commons in June 1940:

> ... **we shall fight** in France, **we shall fight** on the seas and oceans, **we shall fight** with growing confidence and growing strength in the air, we shall defend our Island, whatever the cost may be, **we shall fight** on the beaches, **we shall fight** on the landing grounds, **we shall fight** in the fields and in the streets, **we shall fight** in the hills; we shall never surrender...[6]

This speech was delivered following the successful Dunkirk evacuation of the British troops that had been defending France against Germany during World War 2. The use of epanaphora ensures that key ideas are emphasised. The repetition not only makes the phrase memorable, it adds rhythm to the spoken version. In his speech, Churchill describes honourable defeat, but by using the rhetorical figure, he was able to deliver two messages - that "**we shall fight**" but also that they may very well lose. His successful use of epanaphora ensured that "we shall fight" was the memorable message, while the other went relatively unnoticed (Forsyth, 2014). Epanaphora is one of the rhetorical figures which will be looked at in further detail in this work.

1.1.2 Machine translation

Machine translation (henceforth referred to as MT) involves using computational methods to translate a text between different natural languages (i.e. from French to Spanish). MT has a long history and is deemed to be one of the more challenging areas of AI to solve. Advances in MT have led to increased automation in the field and it is used globally on a large scale basis

[6]Full speech available at `https://winstonchurchill.org/`, courtesy of the Churchill Centre

every day (Way, 2018). However, there is still a requirement for human translators, particularly in the post-editing of machine translated text to ensure it meets the required standard. A human post-editor is an integral part of the MT process and this is not expected to change in the foreseeable future.

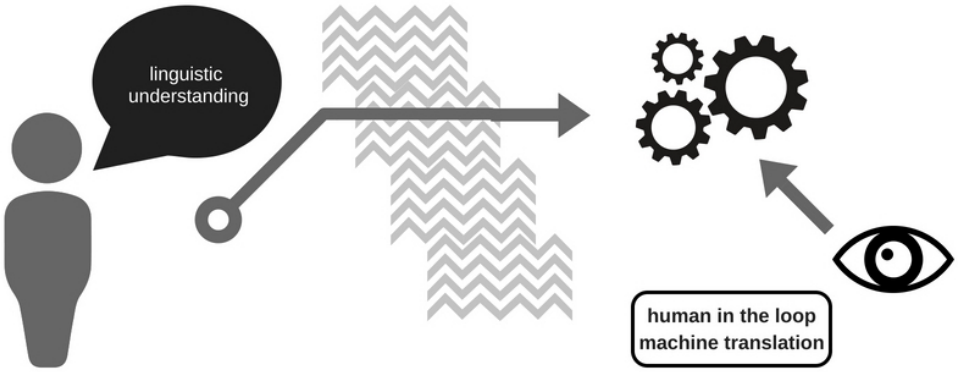

FIGURE 1.1: Proposed method overview

One of the challenges facing natural language researchers and developers is the difficulty in getting their systems to understand and process natural, complex human language such as rhetorical figures. There is potential for a formalised method (depicted in Figure 1.1) which takes linguistic understanding of rhetorical figures and transmits this expertise in a way that could then be utilised in an MT workflow.

The overall aim of this PhD research is propose a method which can bridge a gap between disciplines by transferring domain specific understanding across disciplinary boundaries. The method will be applied to a use case in which linguistic understanding of rhetorical figures is taken and shared with MT practitioners in a way that can be implemented in an MT workflow or system.

1.2 Motivation

> In making a speech one must study three points:
> first, the means of producing persuasion; second, the language; third the proper arrangement of the various parts of the speech.
>
> *Rhetoric*
> *Aristotle*

Despite the rise in interdisciplinary research as highlighted by Larivière and Gingras (2014), McCarty and Deegan (2016) and Van Noorden (2015), and the fact that collaboration, communication and interaction are so central to Digital Humanities (DH) research, Griffin and Hayler (2018) suggest that collaboration is "still under-discussed in the field". They found that when contributors to both of their volumes on research methods in DH in 2016 (Griffin and Hayler, 2016, Hayler and Griffin, 2016) were asked about collaborative research processes, they "were largely met with silence" (Griffin and Hayler, 2018). They observe that despite the rise in and prevalence of collaborative, cross-disciplinary research and projects, those who do such research are reluctant to discuss the collaborative aspects of their work.

Rhetoric is "the study of effective speaking and writing" (*Silva Rhetoricae: The Forest of Rhetoric*). It is commonly referred to as the art of persuasion. Rhetoric is a field of study which has had a "long and vigorous history" (*Silva Rhetoricae: The Forest of Rhetoric*) since its origins in Ancient Greece. The term rhetoric incorporates many different aspects of persuasion such as voice, gesture and logic. This study is concerned with rhetorical figures (which are sometimes referred to as rhetorical devices, figures of rhetoric or figures of speech, but for the sake of clarity will only be referred to as rhetorical figures throughout this work).

Rhetorical figures are defined by Forsyth (2014) as "techniques for making a single phrase striking and memorable". The use of these figures can be observed in real human speech, in arguably all of its forms. Brands use subliminal messages in the language of their advertisements to sell their products

or services. Rhetoric is used both consciously and instinctively in the language of news and media, film, in daily life and by politicians to convince or persuade their electorate. Famous political speeches have been analysed in depth for their use of rhetorical figures and their underlying messages. Since Ancient Greece, public speakers have practised the art of rhetoric and it is an art which has endured and can be observed in political speeches right up to the present day (Martin, 2016).

As long as there has been migration and travel, the necessity to communicate and transmit from one language to another has existed. Translation has been documented as far back as ancient times. In 1799, archaeologists discovered the Rosetta Stone which had been inscribed with three versions of a decree in Egypt in 196 BC. The inscriptions were in two languages and three scripts - Greek, Demotic (Egyptian) and Hieroglyphic (Egyptian). The stone provided the key to the meaning of ancient Egyptian hieroglyphic writing systems for the first time, but it took over twenty years to decode the inscriptions fully (Parkinson, 2005).

MT also has its origins in code-breaking and cryptography. American scientist Warren Weaver is attributed with sowing the seeds for automated translation with this quote:

> One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'. (1947)

Despite a 1966 report by the Automatic Language Processing Advisory Committee (ALPAC) which stated that MT "serves no useful purpose without post-editing, and that with post-editing the overall process is slow and probably uneconomical"[7], research into MT continued and the field has evolved

[7]Language and Machines. Computers in Translation and Linguistics: http://www.mt-archive.info/ALPAC-1966.pdf

from tedious rule based approaches to statistical MT[8] and most recently neural MT[9]. It is predicted that more hybrid MT approaches will emerge which integrate machine reasoning and human expertise.

According to the European Parliament, there is currently "no automatic translation system which can guarantee the high degree of precision and quality required for EU documentation"[10]. In saying that, another political organisation, The European Commission has successfully been using an automated MT system, called eTranslate[11] (formerly MT@EC), to translate documents like press releases, reports and general communications since 2013. The lack of precision is a well-documented issue in MT research, and in the study of natural languages more widely. MT can successfully translate very structured, repetitive texts such as reports and forecasts. However, more complex texts with complicated, nuanced ideas and language still pose problems for MT systems.

To date, there has been little research into translating rhetorical figures, and the work that exists tends to focus on human translation rather than MT. For example, Smith (2006) focused on the translation of rhetorical figures in advertising. The premise of the study was that the aim of advertising is to attract and retain the attention of consumers, and as "the use of rhetorical figures is calculated to have special effects on potential consumers", it is important to understand how such linguistic devices are dealt with in the translation process. Smith's study focused on manual translation, carried out by humans, but the fundamental theories can be applied to MT. Despite advances in the quality of MT output, and advances such as neural MT systems, there has been little focus on the use of rhetorical figures in MT systems to date. This work aims to fill this gap by drawing on rhetorical figure research which may be applicable in the MT field. This includes (but is not limited to) work by Dubremetz and Nivre (2018), Gawryjolek (2009), Harris et al. (2018), Hromada (2011), Java (2015), and O'Reilly and Paurobally (2010).

[8] A Statistical MT system learns by examining a bilingual corpus and produces a translation based on what it has learned

[9] A neural MT system uses an artificial neural network (a type of framework of machine learning algorithms) to predict the likelihood of the sequence of words in the translation

[10] The profession of translator in the European Parliament: `http://bit.ly/EuroParlTranslator`

[11] What is eTranslation?: `http://bit.ly/WhatIs_eTranslation`

While MT is currently being utilised at a commercial scale for tasks such as communicating updates or information, or to sell or provide a service across different languages, MT technology has not yet reached a standard at which it can be implemented in a politically sensitive setting such as the European Parliament without human post-editing. This research posits that as MT output becomes more humanlike, the interspersion of nuanced language such as rhetorical figures will become more applicable.

The aim of this work is to propose a process for cross-disciplinary knowledge transfer which can be built into an MT workflow in practice. In order to do this, it is imperative to develop an understanding of how rhetorical figures function in political speech. This understanding is codified as part of a collaborative process in order that it can be used in an MT system.

1.3 Research questions

Based on the background information and motivation outlined above, the proposed research questions (RQs) for this study are:

1. How can a method of cross-disciplinary knowledge transfer be formalised (e.g. transferring linguistic understanding to MT)? **(RQ1)**
2. To what extent can such a formalised method support the identification of persuasive rhetorical figures in a corpus; and to transfer linguistic understanding of them, to support Machine Translation (MT)? **(RQ2)**

The Oxford Dictionary defines a method as "a particular procedure for accomplishing or approaching something, especially a systematic or established one" while it defines a process as "a series of actions or steps taken in order to achieve a particular end". In the course of this thesis, a method for cross-disciplinary knowledge transfer will be formalised. Formalised, in this context, means that the method will be represented in a process, a set of steps, which can be clearly understood and then subsequently enacted.

1.4 Aims and objectives

In order to address the research questions, the objectives of the research (ROs) will be to:

1. perform a literature review of the areas of interdisciplinary knowledge transfer (with particular reference to digital humanities), rhetorical figures in political speech, Machine Translation and post-editing processes **(RO1)**
2. formalise a process for cross-disciplinary knowledge transfer **(RO2)**
3. present a use case for identifying rhetorical figures from a corpus and transferring the associated linguistic understanding to researchers in the MT domain **(RO3)**
4. assemble a corpus of political speech and a bank of annotated rhetorical figure exemplars in context **(RO4)**

1.5 Methodology

In order to address the research questions, aims and objectives, Chapter 2 reviews literature relating to digital humanities, cross-disciplinary communication and knowledge transfer to ascertain the status of the current research landscape and to identify key directions being explored by research communities. Chapter 4 surveys literature in the area of machine translation in order to highlight the gap and opportunity which exists in terms of integrating linguistic nuance in machine translation systems and processes.

From the literature review, a formalised process for cross-disciplinary knowledge transfer is developed. Philosophies and approaches to knowledge transfer are explored. The process by its nature is interdisciplinary. Therefore it requires collaboration between a minimum of two domains (disciplines). Expertise and understanding from one domain is formalised and made available in a shared location accessible to experts from another domain.

To achieve this, this study adopts a mixed methods approach. Quantitative linguistic analysis is used to derive a set of rhetorical figures that are commonly used in persuasive political speech. These are then codified in XML markup to make them machine-readable, and thus primed for integration in a machine learning process such as machine translation. Qualitative evaluation of the knowledge transfer process is carried out by conducting semi-structured interviews with expert practitioners from the machine translation domain.

1.6 Contribution

This PhD work is interdisciplinary in its nature and the main contributions lie within the following areas: interdisciplinarity, automatic rhetorical figure identification and Machine Translation (MT).

The primary contribution of this work is:

- a method for exchanging knowledge between disciplines which is realised in a process entitled the relo-KT process

The process of analysing rhetorical figure usage in the corpus and the creation of the markup schema incorporates engagement with MT researchers to understand their requirements. Secondary, but significant nonetheless, contributions of this research are:

- an understanding of how persuasive rhetorical figures might be integrated in an MT workflow
- a repository of political statements delivered in Dáil Éireann in January 2018 and a set of annotated rhetorical figure exemplars in the political context

1.7 Thesis outline

The subsequent chapters of this thesis are organised as follows:

Chapter 2 – Literature Review
Following the introduction in this chapter, Chapter 2 outlines the concepts and background related to interdisciplinary research and digital humanities. It also explores knowledge transfer theory and offers an analytical framework for determining whether cross-disciplinary knowledge transfer has taken place in the course of an interdisciplinary research project. The framework, which is rooted in the literature, is applied to five interdisciplinary DH projects.

Chapter 3 – Process Design
Chapter 3 takes the literature outlined in Chapter 2 and uses it to shape a process (the relo-KT process) which formalises how cross-disciplinary knowledge transfer could take place in interdisciplinary research. The six-step process begins with defining the interdisciplinary question and in it, specifying the domains[12] across which understanding needs to be transferred. Steps 2 and 3 are concerned with data acquisition, and the extraction of knowledge from the data as tangible examples. Step 4 centres on the codification of the knowledge extracted in the preceding step while Step 5 involves collaborating with experts from the receiving domain to share the codified knowledge. This is an iterative process in which the interviewees responses can influence the knowledge codification step. The final step of the relo-KT process is to make the data available in an open and sustainable repository.

Chapter 4 – Rhetorical Figures and Machine Translation
Chapter 4 presents additional literature, this time relating to persuasive elements of speech i.e. rhetorical figures, particularly figures of repetition. In this chapter, a history of machine translation is outlined, before exploring how machine translation and rhetorical figures interact in the related literature.

[12]The Oxford Dictionary defines domain as "a specified sphere of activity or knowledge": https://en.oxforddictionaries.com/definition/domain

Chapter 5 – Process Application
In Chapter 5, the relo-KT process is enacted via a use case which features the linguistics and machine translation domains. Chapter 5 addresses RQ2 in that it demonstrates concretely the extent to which the formalised relo-KT process can support the identification of persuasive rhetorical figures from a corpus, and transfer the linguistic understanding of them to the MT domain.

Chapter 6 – Analysis and discussion
Chapter 6 provides an analysis and discussion of the main findings of the research. It is organised in two main sections. In the first, the findings of Chapter 5's use case which explored how the relo-KT process might be enacted to transfer knowledge from the linguistics domain to the MT domain. This is followed by a critical analysis of the overall findings related to the relo-KT process.

Chapter 7 – Conclusion
In this chapter, this study's contribution to knowledge are discussed and the research questions are revisited with reference to the findings. The successes and limitations of the work are presented and suggestions for future work are made.

Chapter 2

Literature Review

2.1 Introduction

This chapter introduces the concept of cross-disciplinarity, and examines how interdisciplinarity is approached, and the challenges which arise when undertaking interdisciplinary research. A set of knowledge transfer techniques is presented, and from these an analytical framework is developed to understand how cross-disciplinary knowledge transfer is carried out in the course of an interdisciplinary research project. In the final section of the chapter, the analytical framework is applied to five interdisciplinary research projects to establish whether cross-disciplinary knowledge transfer occurred during the project.

2.2 Interdisciplinarity

> It takes two flints to make a fire.
>
> *Louisa May Alcott*
> *Little Women (1869)*

Interdisciplinarity, defined as "communication and collaboration across academic disciplines" by Jacobs and Frickel (2009), is “widely practised and theorised” (Franks et al., 2007) across many subject fields and domains. Interdisciplinary research has a long and somewhat chequered past, which has seen it come in and out of fashion numerous times over the centuries (Ledford, 2015). In ’History of Disciplines and Interdisciplinarity’, Szostak notes that

"there have been distinct 'subjects' for thousands of years". Indeed, Plato and Aristotle considered the idea of dividing academic work into discrete categories. As scientific scholarship began to grow, there was soon too much information for one person alone to be an expert in (Ledford, 2015). McKeon (1994) cites the Enlightenment[1] as "the birthplace of modern disciplines", and claims that "it is also the birthplace of modern interdisciplinary studies". It was during the 1800s that the modern disciplinary system we understand started to develop.

The academic disciplines which were established in the nineteenth century "remain the principal organisational unit for the production and diffusion of knowledge" (Weingart, 2010). During the late nineteenth century and the first half of the twentieth century, new categories of knowledge formed and fields such as social sciences emerged. The current push towards interdisciplinary studies started in the 1970s and has been growing since (Ledford, 2015), with many universities establishing large interdisciplinary research centres (Biancani et al., 2018). Many funding programmes, such as the European Commission's Horizon 2020[2], have a requirement for interdisciplinary research (Pedersen, 2016; Pray, 2002). It is well-documented that finding solutions for many of the world's most complex problems (e.g. climate change, water management, information privacy) requires a combination of expertise from varied disciplinary backgrounds (Rylance, 2015).

However, it is the very complexity of such problems which can create a major challenge for the stakeholders involved. Pedersen (2016) cites obesity studies as an example of research which spans multiple disciplines including genetics, metabolism, neuroscience, psychology, ethics, economics, political science and regulation. Attempting to harmonise researchers from such a wide variety of disciplinary backgrounds has the potential to create obstacles. It is worth noting that some disciplines are naturally "more inclined towards interdisciplinary collaboration than others, but it is rare to find researchers today who have no engagement with scholars outside their own speciality" (Pedersen, 2016).

[1]The Enlightenment, also known as the Age of Reason, was a philosophical movement that took place primarily in Europe and, later, in North America, during the late 17th and early 18th century. Its participants were illuminating human intellect and culture after the "dark" Middle Ages

[2]Horizon 2020: `https://ec.europa.eu/programmes/horizon2020/`

2.3 Cross-disciplinarity

Cross-disciplinarity has been described as "an umbrella term" for inter-, trans- and multi- disciplinary studies (Blackmore and Nesbitt, 2008). In order to define cross-disciplinarity, it is necessary to first explore the concepts of *inter-*, *trans-* and *multi-* disciplinary activities as there is a tendency to use these terms synonymously and while each term refers "to the involvement of multiple disciplines to varying degrees on the same continuum" (Choi and Pak, 2006), there are certain differences between them which will be teased out in this section.

Interdisciplinary research is defined by Klein and Newell (1997) as:

> a process of answering a question, solving a problem, or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline or profession... [It] draws on disciplinary perspectives and integrates their insights through construction of a more comprehensive perspective

Choi and Pak (2006) suggest that "interdisciplinarity analyses, synthesises and harmonises links between disciplines into a coordinated and coherent whole", and there is an emphasis on the interaction between disciplines within academia. A prime example of an interdisciplinary project is the Human Genome Project: "an international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings" (*An Overview of the Human Genome Project* 2016). This comprehensive genome mapping project was a collaboration of not only geneticists and scientists, but also experts from the fields of ethics and law.

In contrast, **transdisciplinary** research has the same qualities as interdisciplinary research, but in addition to integrating insights from academic disciplines, it also integrates non-academic insights (*Defining "Transdisciplinary"*). In transdisciplinary research, researchers from different disciplines, other stakeholders and non-researchers transcend "the disciplinary boundaries to look at the dynamics of whole systems in a holistic way" (Choi and Pak, 2006). A transdisciplinary approach is often taken when it comes to managing patient health, as seen in the pain management study by Gordon et al. (2014) in which "clinicians are enabled to implement a unified, holistic, and

integrated treatment plan with all members of the team responsible for the same patient-centered goals".

Like interdisciplinarity and transdisciplinarity, **multidisciplinary** research also draws upon insights from two or more disciplines. However, "unlike interdisciplinary and transdisciplinary activities, multidisciplinarity simply juxtaposes these insights and does not attempt to integrate them" (*Defining "Multidisciplinary" and "Cross-Disciplinary"*). Choi and Pak (2006) describe multidisciplinarity as "the most basic level of involvement" in the sense that different disciplines work on the same problem in parallel, but "without challenging their disciplinary boundaries". Multidisciplinary teams are often found in healthcare or in schools where students require additional support for their educational needs.

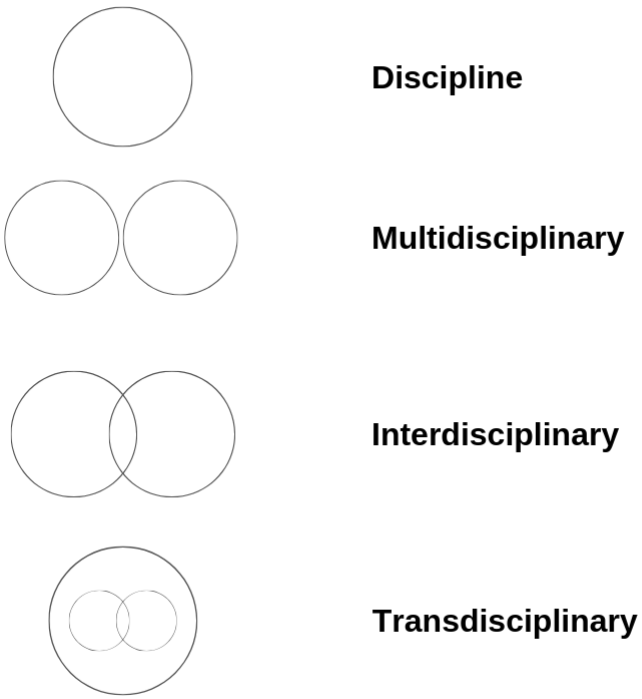

FIGURE 2.1: Visual representation of disciplinarity types

Throughout this research, the terms interdisciplinary and cross-disciplinary will be used. If reference is made to interdisiciplinary or interdisciplinarity, it is Choi and Pak (2006)'s definition that should be borne in mind:

that "interdisciplinarity analyses, synthesises and harmonises links between disciplines into a coordinated and coherent whole". When the term cross-disciplinary is used, it is Szostak's definition that should be kept in mind: that "cross-disciplinary is a general term used to refer to any activity that involves two or more academic disciplines" (*Defining "Multidisciplinary" and "Cross-Disciplinary"*).

2.4 Knowledge transfer

The OECD's[3] Science, Technology and Industry Outlook for 1996 focused on 'the knowledge-based economy'. Knowledge-based economies are those which "are directly based on the production, distribution and use of knowledge and information" (OECD, 1996). The report's aim was to address the growing recognition that knowledge "as embodied in human beings and in technology, has always been central to economic development". According to the report, "knowledge distribution through formal and informal networks is essential to economic performance". With the increasing codification of knowledge for transmission through computer and communications networks, there was also a requirement for tacit knowledge transfer on both an individual and organisational level. This tacit knowledge includes "the skills to use and adapt codified knowledge".

The knowledge-based economy concept has given rise to other related concepts such as knowledge management (Godin, 2006). Knowledge management (KM) is the efficient handling of information within an organisation. In 2000, the OECD published an indepth study entitled *Knowledge Management in the Learning Society*. In it, the organisation acknowledges that despite knowledge being one of the the core elements that drive economies, "it remains hard to understand, measure or systematise its contribution" and that "our knowledge of how knowledge is created, transferred and used remains partial, superficial and partitioned in various scientific disciplines, with the result that the basic concepts of knowledge and learning are defined and interpreted in different ways" (Nelson, 2000). By compiling a record of over 100

[3]The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental economic organisation with 36 member countries. It was founded in 1961 with the aim to stimulate economic progress and world trade. http://www.oecd.org/about/

applied definitions of the term 'knowledge management', Girard and Girard (2015) demonstrate that the concept of KM is one that is recognised in many different sectors from information technology to engineering to health.

Knowledge transfer (KT) is a term which refers to sharing or disseminating knowledge and providing inputs to problem solving. KT "seeks to organize, create, capture or distribute knowledge and ensure its availability for future users" (Lee, 2010). Hence, it is important for interdisciplinary research.

2.5 Interdisciplinary research

As mentioned above, interdisciplinarity involves the analysis, synthesis and harmonisation of links between disciplines (Choi and Pak, 2006), and has been described by Professor Irwin Feller as a "hot topic" in Pray (2002). Szostak and Gagnon (2013) define interdisciplinarity as research which "involves the integration of insights from multiple disciplines in order to better understand some complex topic that is addressed from different perspectives by different disciplines" but maintain that "such a definition tells us a lot about what we are trying to accomplish, but very little about how we might do so".

The perceived, or what Leahey, Beckman, and Stanko (2017) describe as "expected", benefits of conducting interdisciplinary research are numerous and documented by many including Klein (2010); Rhoten (2004); Rylance (2015); and Sanz-Menéndez, Bordons, and Zulueta (2001).

According to the National Academy of Sciences (2005), and as mentioned in Klein (2010), among the driving forces of interdisciplinary research (IDR) are:

- **IDR benefit 1:** the inherent complexity of nature and society. Complex modern problems such as climate change and resource security are not amenable to single-discipline investigation (Rylance, 2015).
- **IDR benefit 2:** the desire to explore problems and questions that are not confined to a single discipline.

- **IDR benefit 3:** the need to solve societal problems. The term interdisciplinary "is often used to denote implicitly or explicitly the application of multiple disciplines and sectors to societal concerns, which may require not only an intellectual answer but perhaps a policy action or technological strategy" Rhoten (2004).
- **IDR benefit 4:** the power of new technologies. Combining expertise from multiple disciplines can lead to creative, high impact research outputs (Bhavsar, 2017).

While the perceived benefits are lauded and aspirational, successful interdisciplinary research is renowned for being difficult to measure. Despite the growth in numbers of researchers undertaking interdisciplinary research, actually carrying it out can present myriad challenges which are synthesised below (Delgado and Åm, 2018; MacLeod, 2018; Sá, 2008):

- **IDR challenge 1:** cognitive obstacles and methodological challenges including "the opacity of domain specific practices to outsiders, conflicting epistemic values, large conceptual and methodological divides and unstructured task environments" (MacLeod, 2018).
- **IDR challenge 2:** there can be a lack of a common language between researchers from different disciplines (Bhavsar, 2017; Della Chiesa, Christoph, and Hinton, 2009) which can in turn lead to:
  - communication challenges - Thompson (2009) suggests that "effective communication patterns help teams to improve and develop rewarding working relationships".
- **IDR challenge 3:** interdisciplinary work requires a lot of time (Jones, 2010) and "significant effort and learning" to overcome "differences in disciplinary culture" (McMurtry et al., 2012).
- **IDR challenge 4:** while "structures to support interdisciplinarity are beginning to be developed in universities and granting agencies, most university promotion systems are still oriented towards narrow disciplinarity" (McMurtry et al., 2012).
- **IDR challenge 5:** interdisciplinary researchers can struggle for prestige as the quantitative metrics system tends to favour single disciplines

(Rylance, 2015) and there can be ambiguity about whether interdisciplinary activities are recognised for career progression (McMurtry et al., 2012).

The above list outlines five common challenges which can arise when carrying out interdisciplinary research. This list presents a context for the types of interdisciplinary challenges which exist. However, this thesis will focus primarily on IDR challenges 1 and 2. The method, and subsequent process, which will be developed in the course of the work aims to reduce the cognitive obstacles and methodological challenges that face researchers when they are undertaking interdisciplinary research. It also aims to address the lack of common language which can impede communication and collaboration in interdisciplinary research. Challenges 1 and 2 are tangible, and addressing them is an achievable aim in the context of this work.

The method which will be presented in the course of this thesis does not attempt to address the other challenges. For example, this method will not strive to reduce the time interdisciplinary research takes (IDR challenge 3). While efficiency is of course desirable, it not a priority in terms of this work. Challenges 4 and 5 address more institutional or departmental concerns which fall outside the remit of this study.

2.5.1 Interdisciplinary knowledge transfer

One of the tenets of interdisciplinary research is its collaborative ethos. Formalised approaches to knowledge sharing are necessary to ensure seamless exchange of knowledge - both explicit and tacit. Explicit knowledge is knowledge which is easy to articulate and codify. An example of explicit knowledge is the type of information which can be stored in spreadsheets. We are living in a knowledge-driven world in which technology allows us to transmit explicit knowledge rapidly and easily. Tacit knowledge, on the other hand, tends to be difficult to transfer in a codified way. Generally, it is 'grey hairs' knowledge which is intuitive, observed or gained from personal experience.

It is important to briefly differentiate between the terms data, information and knowledge here as there can be a tendency to use these terms synonymously. The data–information–knowledge–wisdom (DIKW) hierarchy (Figure 2.2) is "one of the most fundamental, widely recognized and 'taken-for-granted' models in the information and knowledge literatures" (Rowley, 2007). It represents the relationship between data, information, knowledge and wisdom in a pyramid with data at the base and wisdom at the pinnacle. Rowley's work reviews "popular explicit or implicit articulations" of DIKW from relevant textbooks in the areas of knowledge revolution, information systems and knowledge management. Rowley reviews and presents a number of definitions of each term and these are summarised below.

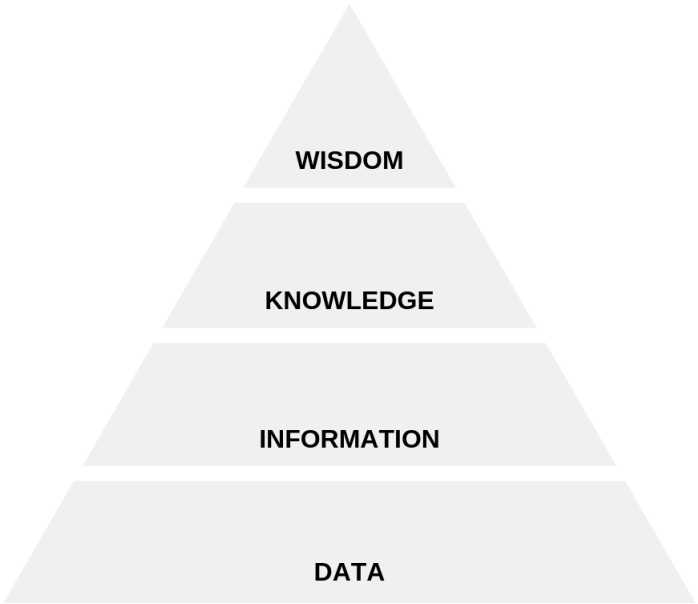

FIGURE 2.2: Data-information-knowledge-wisdom hierarchy as summarised by Rowley (2007)

- **Data** is facts and statistics collected together for reference or analysis[4];
- **Information** is data which has been given a meaning, relevance and purpose (Jashapara, 2004);
- **Knowledge** is the combination of data and information, to which expert opinion, skills, and experience is added. This results in a valuable asset which can be used to aid decision making (definition in Rowley (2007) from Chaffey and White (2010), quoting the European Framework for Knowledge Management[5]);
- **Wisdom** is an evaluated understanding (attributed to Ackoff); the ability to make sound judgements and decisions apparently without thought (Wallace, 2007).

Lundvall and Johnson (1994) divide knowledge into four categories where know-what and know-why (1 and 2) correspond to Rowley's data and information while know-how aligns more closely with knowledge on the hierarchy:

1. **Know-what** refers to knowledge about facts;
2. **Know-why** refers to knowledge about principles and laws of motion in nature, in the human mind and in society;
3. **Know-how** refers to skills;
4. **Know-who** refers to information about who knows what and who knows what to do. Additionally, know-who can encompass the ability to co-operate and communicate with different kinds of people.

This concept of know-how is of particular relevance to this work. Know-how (typically also incorporating elements of know-who) can be likened to tacit knowledge, the type of knowledge which develops with experience and is generally difficult to capture and codify. Johnson, Lorenz, and Lundvall (2002) describe know-how as "the kind of knowledge where information

[4]Oxford Dictionary: https://en.oxforddictionaries.com/definition/data
[5]European Framework for Knowledge Management:
http://enil.ceris.cnr.it/Basili/EnIL/gateway/europe/CEN_KM.htm

technology faces the biggest problems in transforming tacit or non-explicit knowledge into an explicit, codified format". Although Lundvall and Johnson (1994) refer to know-how as skills knowledge, this work is concerned with **know**ing **how** persuasive language is arranged in order to begin to codify it for transfer across disciplines.

The OECD's report on knowledge management highlights the importance of codifying knowledge in order to share it more easily (OECD, 2000) as did Cowan and Foray (1997) who stated that "the process by which knowledge or information evolves and spreads through the economy involves changing its nature between tacit and codified forms". While a lot of the literature in this area focuses on the economics of knowledge, the fundamental theories of knowledge transmission can be seen as universal. The standardisation or codification of knowledge in order for it to be transferred and reused is something which can be beneficial to multiple sectors, ranging from technology to education, to science, to retail.

Codified tacit knowledge is embedded in every aspect of our lives. It is maintenance manuals, recipes, knitting patterns and project documentation. Through codification, knowledge can be conveyed and quantified.

2.5.2 Analytical frameworks for interdisciplinary research

There are a wide range of methodologies for interdisciplinary research. Griffin and Hayler (2018) found that silence persists when it comes to discussing methodological approaches to interdisciplinary research projects while Rawlings et al. (2015) observe that "knowledge flows remain an underanalysed aspect of academic research". Regardless, work has been carried out to develop frameworks that can be applied when undertaking interdisciplinary research. One of the most relevant frameworks for interdisciplinary research is Szostak (2002)'s twelve-step process, which is presented in this section alongside the PMI (2015)'s knowledge transfer life cycle.

Szostak's twelve-step process for interdisciplinarity

In 'How to Do Interdisciplinarity: Integrating the Debate' Szostak (2002) develops a twelve-step process for interdisciplinary research. In it, he differentiates between individual researchers and "communities of interdisciplinary researchers" and notes that while "individual researchers cannot be expected to follow all of these steps in every research project, the process alerts them to the dangers of omitting steps", but that communities (i.e. research project teams) "should ensure that all steps are followed". Szostek's work builds on earlier work by Newell, Wentworth, and Sebberson (2001) and Klein, Wentworth, and Sebberson (2001).

Szostak (2002)'s twelve step process for undertaking interdisciplinary research can be summarised as follows:

1. Start with an interdisciplinary question;
2. Identify the key phenomena involved, but also subsidiary phenomena;
3. Ascertain what theories and methods are particularly relevant to the question at hand (be careful not to casually ignore theories and methods that may shed some lesser light on the question);
4. Perform a detailed literature survey;
5. Identify relevant disciplines and disciplinary perspectives;
6. If some relevant phenomena (or links among these), theories, or methods identified in (2) and (3) have received little or no attention in the literature, the researcher should try to perform or encourage the performance of such research;
7. Evaluate the results of previous research;
8. Compare and contrast results from previous disciplinary or interdisciplinary research;
9. Develop a more comprehensive/integrative analysis;
10. Reflect on the results of integration;
11. Test the results of integration;
12. Communicate the results.

While Szostak's twelve steps provide important guidance for those undertaking interdisciplinary research, they do not detail a how-to approach to imparting subject or domain specific expertise and know-how. There is a space in Szostak's twelve steps where knowledge transfer could be integrated in order to provide more holistic guidance on interdisciplinary research. In Szostak's approach, step 9 is where the main interaction between the disciplines occurs - he describes it as the "integration" of disciplines. Szostak makes reference to Bailis (2001) who stresses that "integration may proceed quite differently depending on the question addressed". Szostak follows this by saying that researchers themselves should "ascertain which types of integration are most important for particular questions" and also mentions that "a common vocabulary would be invaluable" at this stage (lack of a common language was earlier outlined as IDR challenge 2 in section 2.5).

The Project Management Institute (PMI)'s knowledge transfer life cycle

The Project Management Institute PMI is a worldwide nonprofit professional organisation for project management. According to its 2017 annual report[6], the PMI has over 500,000 members in 207 countries, making it "one of the world's largest membership-based professional societies". It awards certification for project management professionals which is recognised globally[7] and it also provides professional training and development for its members[8].

In 2015, the PMI's 'Capturing the Value of Project Management Through Knowledge Transfer' publication outlined a life cycle for knowledge transfer. One of the aims of the report was to demonstrate that "being good at knowledge transfer improves project outcomes" (PMI, 2015). The life cycle (Figure 2.3) presents the KT process as a cyclical one which does not necessarily have a defined beginning or end point, indicating that KT should be seen as an ongoing, iterative endeavour.

[6]PMI 2017 Annual Report: https://www.pmi.org/annual-report-2017/at-a-glance
[7]PMI Certifications: https://www.pmi.org/certifications
[8]PMI Training & Development: https://www.pmi.org/learning/training-development

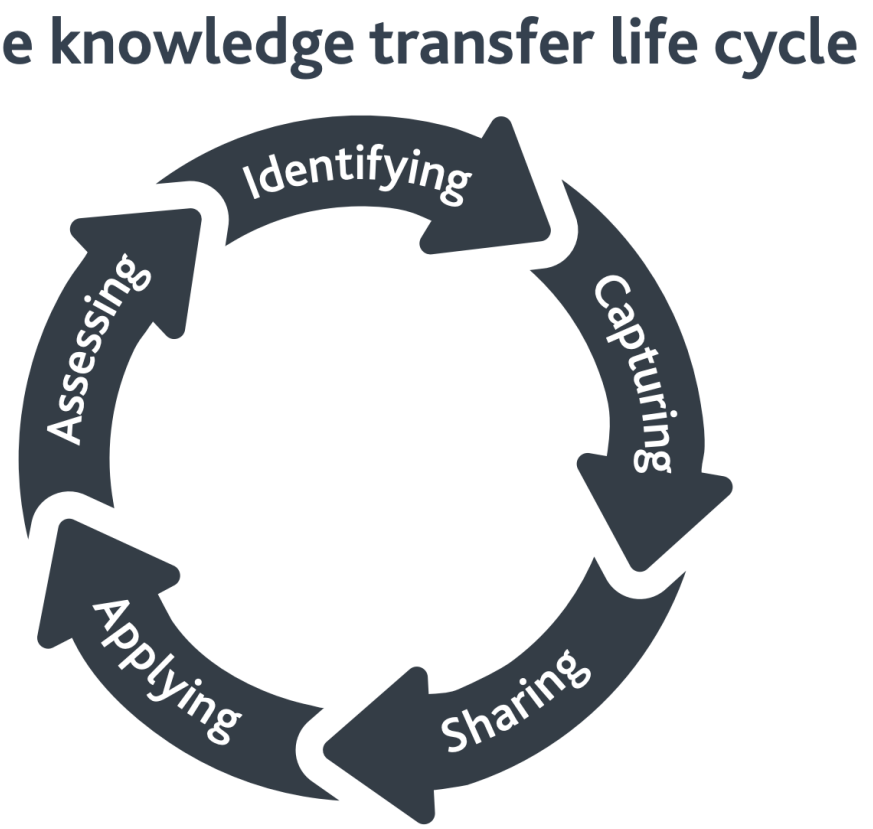

FIGURE 2.3: Project Management Institute's Knowledge Transfer Life Cycle

The Knowledge Transfer Life Cycle (PMI, 2015) captures five key steps which are required in the KT process. These steps represent a systematic visualisation of the knowledge transfer process:

1. **Identifying:** determine the knowledge which needs to be transferred;
2. **Capturing:** accumulate the essential knowledge that needs to be transferred;
3. **Sharing:** establish methods for transferring the knowledge;
4. **Applying:** use the knowledge that is transferred;
5. **Assessing:** evaluate the benefits of the knowledge that is transferred.

2.6 Interdisciplinarity and digital humanities

In the specific case of digital humanities (DH), often described as research which takes place at the intersection or nexus of humanities and computing, Hayles (2012) suggests that collaboration should be the "rule rather than the exception" and **edmond_toward_2015** state that "digital humanities is highly interdisciplinary and highly collaborative". Griffin and Hayler (2018) point out that collaboration has "gradually come to be a demand, if not an explicit necessity, for humanities scholars".

In 2003, Robertson, Martin, and Singer (2003) found that a silence persisted "about the methods of interdisciplinary collaboration" although many studies into interdisciplinarity had been carried out "especially in relation to the creation of new disciplines and institutions". While researchers reported on methods of data collection and analysis, they would "seldom report the methods they employ in the process of achieving interdisciplinary collaboration itself". Six years later, Dykes, Rodgers, and Smyth (2009) called for "a consistent disciplinary framework" for contemporary creative design practice.

Recent literature suggests that despite the emphasis being placed on interdisciplinary research from funding bodies, universities and faculties, there is a disconnect with what happens in research practice. Griffin and Hayler (2018) found that the topic of collaboration in DH is "still under-discussed in the field". They note that when contributors to their two volumes on research methods in DH were asked about collaborative research processes (Griffin and Hayler, 2016; Hayler and Griffin, 2016), they "were largely met with silence" (Griffin and Hayler, 2018).

This indicates that in interdisciplinary projects, it can be difficult to discern how cross-disciplinary domain knowledge is transferred, as this type of information tends not to be documented. Interdisciplinary research project teams may craft bespoke methodologies appropriate to the task at hand and the decisions or processes which led to their crafting can be hidden in minutes of meetings or in discussions which are not made public. For this reason, it can be hard to understand where the interdisciplinary exchange occurred.

There is a long list of DH projects, many of them interdisciplinary, which could be included in a review of cross-disciplinary transfer of domain understanding. Website trawls of DH centres at universities such as Stanford[9], Berkeley[10] and University College London[11] demonstrate the wide variety of interdisciplinary DH research which is being carried out. The European Association for Digital Humanities (EADH) maintains a list of projects undertaken during the previous five years which "contribute meaningfully to DH in Europe"[12]. In December 2018, there were over 200 projects listed. In a similar vein to the DARIAH (Digital Research Infrastructure for the Arts and Humanities) DH course registry[13] (which shares information about higher education DH courses in the field), the Erasmus University Rotterdam and CLARIAH-NL (Common Lab Infrastructure for the Arts and the Humanities) maintains a DH Project Registry[14] which provides an overview and visualisations of over 300 DH projects which have been carried out in the Netherlands since 1989.

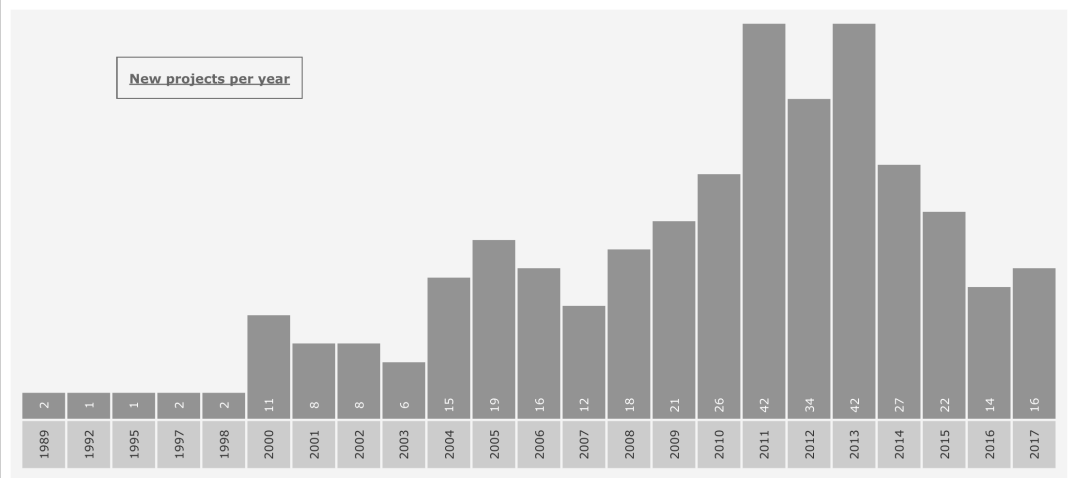

FIGURE 2.4: Projects added to the DH project registry 1989-2017

[9]DH Projects at Stanford: https://digitalhumanities.stanford.edu/projects
[10]DH at Berkeley Projects: https://digitalhumanities.berkeley.edu/projects
[11]UCL Centre for DH Research Projects: https://www.ucl.ac.uk/dh/projects
[12]EADH Projects: https://eadh.org/projects
[13]DH Course Registry: http://bit.ly/DH_registry
[14]DH Projects Registry: http://dh-projectregistry.org

2.6.1 Establishing an analytical framework

Five interdisciplinary DH projects will be analysed in an effort to understand whether, and how, cross-disciplinary knowledge transfer is carried out in the course of an interdisciplinary research project. The preceding sections provide a lens through which domain specific knowledge transfer across disciplines can be analysed. With regard to the research question (RQ1) stated in Chapter 1: 'How can a method of cross-disciplinary knowledge transfer be formalised?', there are a number of key things to consider.

Firstly, there is the **interdisciplinary** aspect as seen in Szostak (2002)'s twelve-step 'How to do Interdisciplinarity'. In order to determine whether the DH projects achieve cross-disciplinary KT, they will be examined keeping Choi and Pak (2006)'s definition in mind - that the project harmonises links and findings from different disciplines into "a coordinated and coherent whole".

Knowledge transfer, as seen in the previous section, can be complex and multifaceted. Conducting interdisciplinary research can also be fraught with challenges, among them the cognitive and methodological obstacles and lack of common language challenges established in section 2.5. In order for tacit knowledge to be transferred from one domain to another, it must first be **codified** in an appropriate way (OECD, 2000; Cowan and Foray, 1997).

Evaluation of interdisciplinary research is a challenging task, primarily due to differing perspectives on what quality means in different disciplines. In Klein (2008)'s literature review of interdisciplinary and transdisciplinary research evaluation, she states that evaluation "remains one of the least-understood aspects" of interdisciplinary research while Huutoniemi (2010) reflects on challenges such as achieving balance between different epistemic viewpoints and who should be included in the evaluation process. In terms of KT evaluation, the PMI's Knowledge Transfer Life Cycle addresses the need to evaluate the benefits of the knowledge transfer (Assessing). Research on knowledge transfer carried out on 2,466 project management practitioners by PMI (2015) found that when it comes to the KT life cycle, "approximately two-thirds of organizations follow the first three steps — identifying knowledge, capturing and retaining knowledge, and making knowledge available" but "few follow the last two steps": applying the transferred knowledge and

assessing the value or benefits of specific knowledge. The report found that the value of **knowledge transfer** is "difficult to measure because it's not always tangible or precise".

Huutoniemi (2010) quotes the OECD (1998) which stated that "highly competent proficiency in a single discipline is the only acceptable basis for interdisciplinary success". Additionally, she quotes a symposium of interdisciplinary experts from Harvard University and the American Association for the Advancement of Science who state that "a basic premise of quality interdisciplinary work is that it satisfies quality standards arising from the disciplines involved". Therefore, excellence within the individual disciplines is an essential element for the quality transfer of tacit knowledge between them. The PMI suggests a cyclical approach to tacit knowledge transfer.

	Derived from	**Relevant aspects to this work**
inter-disciplinarity	Szostak (2002)	Step 9: where interaction between the disciplines occurs: develop a more comprehensive/ integrative analysis
codify knowledge to enable KT	Cowan and Foray (1997) OECD (2000)	In order for tacit knowledge to be transferred from one domain to another, it must first be codified in an appropriate way
evaluation	Huutoniemi (2010) PMI (2015)	Excellence within the individual disciplines Iterative cycle of tacit KT

TABLE 2.1: Analytical framework for determining whether, and how, cross-disciplinary knowledge transfer has been achieved in interdisciplinary research

Table 2.1 outlines the steps of the framework which will be applied to determine whether, and how, cross-disciplinary transfer of domain understanding was carried out in the execution of these projects.

2.6.2 Digital humanities projects

While the list of interdisciplinary DH projects is long and diverse, I have chosen to apply the analytical framework outlined above to five interdisciplinary DH projects to determine how cross-disciplinary transfer of domain understanding was handled. Desk research was carried out to find suitable interdisciplinary DH projects. Each of the five projects were chosen because:

1. they involve collaborative effort between two or more disciplines, and have produced some concrete interdisciplinary output;
2. they are well documented, either on a project website, in peer-reviewed publications, or both.

In the descriptions which follow, I present a brief overview of each project before analysing the particular approaches each one took with regard to cross-disciplinary knowledge transfer. To keep to within the analytical framework (Table 2.1), each project was examined with the following questions in mind:

1. is the project interdisciplinary - does it merge knowledge, insight and methods from two or more disciplines into "a coordinated and coherent whole"?
2. is project-specific knowledge codified in a formal way?
3. is any evaluation carried out on the project?
4. is project-related knowledge transfer between the disciplinary stakeholders documented?
    - if yes, how does it do this?

- even if it is not explicitly documented, is there evidence of KT?

The criteria were chosen to examine each project in terms of its cross-disciplinary knowledge transfer. It was a challenge to find one project which either a) explicitly set out to do interdisciplinary transfer, or b) which explicitly documents the process of interdisciplinary transfer. This suggests that the modes of cross-disciplinary knowledge transfer tend not to be discussed explicitly when it comes to interdisciplinary DH projects, thus leading to the silence which Griffin and Hayler (2018) address.

CULTURA

CULTURA is a personalised virtual research environment which was developed to address a challenge which faces the curators and providers of digital cultural heritage: namely how to "instigate, increase and enhance engagement with digital humanities collections" (*CULTURA (Cultivating Understanding and Research through Adaptivity)*). The environment itself is corpus agnostic. It contains "a suite of services, including personalisation, annotation, and recommendation, providing necessary supports and features for a diverse range of professional researchers" (Conlan et al., 2014).

Deep Maps: West Cork Coastal Cultures

Deep Maps: West Cork Coastal Cultures is a project which explores "the rich maritime environment that is found along the arc of Cork's Roaring Water Bay, from Timoleague to Bantry Bay, as it is shaped by sea and land and imagined in literature and art" (*Home*). The project was a collaboration between University College Cork's School of English and School of Biological, Earth and Environmental Sciences. The project's website highlights the interconnection of personal, cultural, historical and scientific knowledge which is associated with coastal environments. The overall aim of the project is to "develop new ways of thinking about place" and it achieves that by bringing together an interdisciplinary team which combined "the research skills of cultural historians with those of marine biologists" (*About the Project*) and added

in DH research skills such as 3D visualisation techniques and geographic information systems (GIS) to create the *Deep Maps: West Cork Coastal Cultures* website which presents contextual data from a wide range of sources.

Industrial Memories

In 2009, the Irish government published the Ryan Report which "detailed the findings of a 9 year investigation into abuse and neglect in Irish industrial schools" (Leavy, Pine, and Keane, 2018). The report is in five volumes and runs to over 500,000 words spread across more than 2500 pages (Leavy, Pine, and Keane, 2017). Due to the size of the report, Leavy, Pine, and Keane (2018) claim that it "remains largely unread". *Industrial Memories* was a collaborative project undertaken between the School of English, Drama and Film at University College Dublin UCD and the UCD Insight Centre (a data analytics research centre). The interdisciplinary team developed "a web-based platform where the narrative form of the Ryan Report is deconstructed and key information extracted" (Leavy, Pine, and Keane, 2018). Using text analysis and other methods of distant reading and visualisation, the Industrial Memories project provides an insight into the otherwise inaccessible Ryan Report (due to its structure and size) (Pine, Leavy, and Keane, 2017).

Old Weather

The *Old Weather* project is a crowdsourced project in which "volunteers explore, mark, and transcribe historic ship's logs from the 19th and early 20th centuries" (*Old Weather - About*). The logs were originally completed by mariners and scientists aboard ships and span a 150 year period. The handwriting is idiosyncratic and thus requires human intervention to produce two types of results:

1. weather observations;
2. historical records.

Participants in the *Old Weather* project are "helping advance research in multiple fields" when they annotate and transcribe the ships' logbooks. The project website claims that data on past weather and sea-ice conditions which is produced is "vital for climate scientists, while historians value knowing

about the course of a voyage and the events that transpired" (*Old Weather - About*).

Transcribe Bentham

Transcribe Bentham is an online crowdsourced transcription initiative which has been underway at University College London (UCL) since 2010. At the latest count[15], the public has transcribed over 20,000 original and unstudied manuscript papers written by philosopher and reformer Jeremy Bentham (1748-1832) (*About Us - Transcribe Bentham*).

2.6.3 Applying the analytical framework

The analytical framework outlined in Table 2.1 provides a lens through which the DH projects can be examined to determine whether cross-disciplinary KT was achieved. The observations are presented in this section.

Interdisciplinarity

Each of the projects described in section 2.6.2 can be defined as interdisciplinary because in each, researchers from two or more disciplinary backgrounds collaborated to create an interdisciplinary artefact.

Codify knowledge to enable KT

Various approaches were taken to codify the knowledge or understanding produced in the projects. This is due to the varying nature of the input data acquired by each project. This ranged from textual corpora in the case of *CULTURA*, *Industrial Memories* and *Transcribe Bentham* and tabular and textual data of "millions of weather, ocean, and sea-ice observations recorded by

[15]This information was noted on the *Transcribe Bentham* in February 2019

mariners and scientists over the past 150 years" gathered by the *Old Weather* project (*Old Weather - About*). The *Deep Maps* project brings together a wide range of data, including biological/scientific, cultural (e.g. art, literature and folklore) and historical (e.g. official data and personal accounts) spanning a time period from 1700 to present day (Murphy and Power, 2017).

In the projects which used textual corpora, *Transcribe Bentham* uses the TEI-XML markup language which has become a standard for systematically encoding texts while Conlan et al. (2014) describe the entity extraction procedure used to extract and markup entities from historical documents. The *CULTURA* system is corpus agnostic, but it was "designed in such a way to be deployed in different configurations across different collections" (Conlan et al., 2014). One of the data interventions *CULTURA* makes is to decode digitised text into a machine-readable representation. This involves extracting entities from a corpus and applying usable mark-up in the form of XML.

The *Industrial Memories* project, which was also text-based, represented The Ryan Report in a relational database with annotated excerpts comprising 6,839 paragraphs (597,651 words). Each paragraph was considered a unit-of-analysis in the Report and was represented as a database instance. Leavy, Pine, and Keane found that each paragraph tended to focus on a particular topic which meant that they could be categorised and annotated (2017). The paragraph categories were identified using methods such as: automated text classification; rule-based searching and "feature selection based on context-specific semantic lexicons generated from a sample of seed-words using a word embedding algorithm" (Leavy, Pine, and Keane, 2017). Named entities were automatically extracted using NLTK[16].

In *Deep Maps*, the data gathered is represented multi-modally using tools such as the multi-layered interactive Deep Map of West Cork[17] and the story map which focuses on travellers' accounts of West Cork[18]. The project used tools such as ArcGIS[19] and ESRI Story Maps[20] (which are built into ArcGIS) to display the data visually. Tabular data from spreadsheets can be added

[16]The Natural Language Toolkit: https://www.nltk.org/
[17]Deep Map of West Cork: http://bit.ly/DeepMapWestCork
[18]Travellers' Accounts and West Cork: http://bit.ly/TravellersAccountsMap
[19]ArcGIS: https://www.arcgis.com/index.html
[20]ESRI Story Maps: https://storymaps.arcgis.com/en/

to these tools to generate the maps. Finally, once the *Old Weather* data has been transcribed, numerous processes carried out on the data before it is ultimately converted into a standard format called the International Marine Meteorological Archive (IMMA) format that can be easily used by professional scientists in major research projects (Spencer, 2018).

Evaluation

In terms of evaluation, not all of the projects have documented their evaluative processes. The *Transcribe Bentham* project uses human evaluation to check and approve the accuracy of the transcribed words and associated tags (Causer et al., 2018). In terms of data processing for the *Old Weather* project, the transcribed data for both latitude and longitude positions and weather observations are analysed and checked for quality control by a member of the science team before being added to a large database (Spencer, 2018).

When developing the *CULTURA* system, the accuracy of the entity extraction from historical documents was evaluated and measured by "comparing automatic pipelined output to manual mark-up, and has been shown to be high" (Conlan et al., 2014). The *CULTURA* virtual research environment itself was also tested and deployed as outlined by Steiner et al. (2014).

Evaluation procedures for the *Deep Maps* and *Industrial Memories* projects were not found within the online documentation analysed in this study.

Cross-disciplinary KT

While cross-disciplinary KT was not an explicit aim of the *CULTURA* project, it is clear from analysis of the project's publications that the numerous disciplines involved collaborated and worked together to create the *CULTURA* system which weaves together humanities data and insight with computer

science technologies and methodological approaches. The result is an adaptive and "interactive user environment which dynamically tailors the investigation, comprehension and enrichment of digital humanities artefacts and collections".

Similarly, interdisciplinary communication and collaboration were clearly used in the development of the Industrial Memories project. Distant reading methods were utilised to draw patterns from the text that would be very difficult (if not impossible) to glean from a close reading of the report's narrative. While there isn't a direct reflection on the interdisciplinary collaborative process, in the three publications related to this work (Leavy, Pine, and Keane, 2017; Pine, Leavy, and Keane, 2017; Leavy, Pine, and Keane, 2018), or on the comprehensive website (`https://industrialmemories.ucd.ie/`), it is clear that a digital edition of this scale could not be completed without cross-disciplinary collaboration and KT.

Causer et al. (2018) carried out an evaluation of the participatory nature of the *Transcribe Bentham* project and found that "it is clearly a complex task to evaluate the efficiencies and economics of cultural heritage crowdsourcing". The paper offers general recommendations for any teams who are undertaking "large-scale crowdsourcing for cultural heritage" but does not cogitate on the collaborative and interdisciplinary efforts which took place over the course of the project.

Interestingly, some of the recommendations Causer et al. present mirror points outlined in section 2.5 in terms of the challenges of carrying out interdisciplinary research in general and suggest that a successful crowdsourced project "requires an ambitious and well thought-through project plan at the very beginning, and ongoing institutional support, commitment, and resources to successfully meet the crowdsourcing programme's goals". Ultimately, Causer et al. (2018) conclude that the *Transcribe Bentham* project is a successful one. While the collaborative and interdisciplinary effort is not explicitly discussed, it has obviously taken place over the project's lifespan.

Project	Inter-disciplinarity	Codify knowledge to enable KT	Evaluation	Cross-disciplinary KT
CULTURA	YES	XML	comparison of manual & automated output questionnaire semi-structured moderated discussion log data	not documented
Deep Maps	YES	3D visualisation GIS	not documented	not documented
Industrial Memories	YES	automated text classification	not documented	not documented
Old Weather	YES	International Marine Meteorological Archive (IMMA)	human intervention	not documented
Transcribe Bentham	YES	TEI-XML	human intervention	not documented

TABLE 2.2: Survey of DH projects' cross-disciplinary knowledge transfer

Overview of projects

From the project descriptions, Table 2.2 was generated to present a view of how visible the domain understanding (knowledge) transfer was in each project when analysed using the framework presented in Table 2.1. The columns in Table 2.2 represent tangible artefacts of domain understanding: whether knowledge codification of any sort was performed, whether the project utilises a data repository to ensure access and longevity even after the project's life cycle has ended, and whether evaluation of how well domain understanding transfer was performed.

All five projects are interdisciplinary. Each one uses at least one method of knowledge codification. The methods chosen range from discipline specific (e.g. *Old Weather*'s use of the IMMA annotation format) or uses a de facto standard (*Transcribe Bentham*'s use of TEI-XML and *CULTURA*'s use of XML).

Evaluation methods vary which reflects the PMI (2015)'s finding that few organisations follow the step which involves "assessing the value or benefits of specific knowledge". Only Steiner et al. (2014) presents an evaluation model which has "high potential for reuse in other research environments".

In terms of cross-disciplinary KT, none of the projects expressly documented their collaborative or KT processes. However, these endeavours were not absent, they were merely latent because disclosing the collaborative methods used was not an aim for any of the projects. This reflects some of the literature outlined earlier in this chapter (PMI, 2015; Griffin and Hayler, 2018).

2.7 Summary

In this chapter, the concepts of cross-disciplinarity and knowledge transfer were examined. Undoubtedly, interdisciplinary research is a worthwhile and valuable approach to scholarship and is potentially a key to solving some of the world's most complex problems. One of the challenges which interdisciplinary collaboration presents is that communication can break down due to the lack of a common language between participants. As a result, researchers

may feel demotivated. This can result in knowledge not being shared and projects may fail (in the sense that objectives are not achieved).

Despite formalised processes for interdisciplinary collaboration (Szostak, 2002; PMI, 2015), studies have found that researchers involved with collaborative DH projects tend not to discuss and divulge their cross-disciplinary knowledge sharing methodologies which indicates that there is a gap in the area that needs to be addressed (Griffin and Hayler, 2018). In order to examine this further, an analytical framework was developed to examine the extent to which cross-disciplinary KT occurs in interdisciplinary research projects. When the analytical framework was applied to five DH projects, it demonstrated that while collaborative practices don't tend to be documented, they clearly occurred in latent ways over the course of the projects.

The insights gleaned from the literature in this chapter will be utilised in developing the relo-KT process for cross-disciplinary KT which is presented in Chapter 3.

Chapter 3

relo-KT Process Design

3.1 Introduction

Chapter 2 outlined the literature related to interdisciplinarity and DH. One of the sections presented five interdisciplinary DH projects and explored how domain understanding (the expertise or know-how) associated with a specific subject area can be transferred. This chapter presents the design of a process to facilitate cross-disciplinary knowledge transfer. It is envisaged that the process described in this chapter fits into and complements a model such as Szostak's 12 steps for undertaking interdisciplinary research, as outlined in section 2.5.2.

The aim of this process is to facilitate a tangible transfer of domain knowledge. This means going beyond a theoretical or conceptual mode, to an example based model. This chapter will explore how to take tacit knowledge or expertise on a subject and represent that knowledge in codified examples so that it can be taken up by a practitioner or researcher from another domain.

3.2 Design requirements

As Chapter 2 demonstrated, there are different ways of carrying out knowledge transfer, be it the transfer of data and information or more complex know-how, also known as tacit knowledge. Chapter 2 also revealed an opportunity to develop a process for cross-disciplinary KT which could be integrated into an interdisciplinary research process.

Based on the literature by Lundvall and Johnson (1994), Johnson, Lorenz, and Lundvall (2002), and Cowan and Foray (1997), it is clear that a cross-disciplinary knowledge transfer process requires taking tacit knowledge and codifying it in order to be able to facilitate transfer.

3.3 Design description

This chapter will focus on the **relo-KT** process which has been developed to facilitate cross-disciplinary KT. The relo-KT process (pronounced 'relocate') gets its name from a play on the word 'relocate' meaning "move to a new place" and the KT of Knowledge Transfer. Essentially, the process involves sharing knowledge from one domain (D.1), and establishing it in a shared repository which is accessible to the other (D.2).

3.3.1 Overview

To provide a précis of the process, I argue that with an appropriate corpus, domain knowledge or expertise can be extracted and codified to provide concrete examples of domain understanding. These examples become the means for transferring the theoretical knowledge of one domain (D.1) to another (D.2). Drawing on the literature from Chapter 2, I have outlined a number of steps which need to be considered in order to successfully transfer knowledge across disciplines:

These steps are depicted in figure 3.1. The numbering of the steps reflects the order of the process as it was carried out in the course of this study. However, these steps are not prescribed to be followed in this order. Depending on the nature of the inquiry, a different order may be required, or the process may need to iterate a number of times.

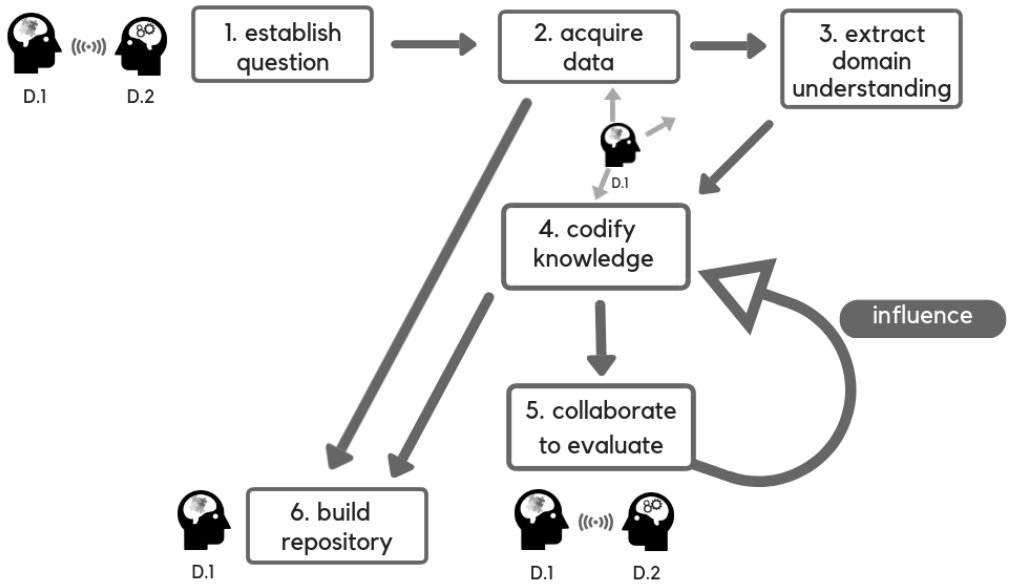

FIGURE 3.1: The relo-KT process for cross-disciplinary knowledge transfer

3.3.2 The relo-KT process

Figure 3.1 provides an overview of the relo-KT process for cross-disciplinary knowledge transfer . This section will provide a more detailed breakdown of each component of process and describe how the components are linked and the interaction between them.

Step 1: Question

The first step of Szostak (2002)'s 12 step process is "Start with an interdisciplinary question". Szostak highlights the necessity of taking time to develop and identify a question or questions which will guide the research process. Tress et al. (2003) claim that in order for research to be interdisciplinary, "the research question is defined jointly and the answer to the research questions derives from an integration of disciplinary knowledge".

Creswell (2014) outlines the considerations required for developing research questions, depending on whether the research is qualitative, quantitative or mixed methods research. Mixed methods research "resides in the middle" of the continuum which sees qualitative research at one end and quantitative at the other (Creswell, 2014).

According to the National Academy of Sciences and Engineering "interdisciplinary research is a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice".

Step 2: Acquire data

In order to answer the research question from step 1, it is necessary to acquire data. As seen in Chapter 2, data can be described as factual observations, measurements and statistics. Information can be defined as data which has been given some kind of meaning or purpose (Rowley, 2007) and assembled in a systematic way. Data or information can come in a variety of forms depending on the research area. As mentioned in Chapter 1, the main area of focus in this study is how linguistic understanding can be transferred to the MT domain. For this reason, the research methods presented here will have a distinct linguistic focus.

In 'Using Corpora for Discourse Analysis', Baker (2006) presents methodological techniques and approaches to using corpus-based research, and notes that while corpus-based investigation is a quantitative exercise, it also "involves a great deal of human choice at every stage".

Creswell (2014) outlines data collection methods for qualitative and quantitative research. These include:

- **Qualitative:** interviews (e.g. in person one-on-one or focus group), observations (e.g. researcher observes a participant), documents (e.g. public documents such as meeting minutes or private documents such

as journals or diaries), audio-visual materials (e.g. photographs or art objects)

- **Quantitative:** surveys, experiments, empirical observations and measures

According to Creswell (2014), by drawing on both qualitative and quantitative approaches, mixed methods research "provides a sophisticated, complex approach to research that appeals to those on the forefront of new research procedures" at a practical level, while at a procedural level, "it is a useful strategy to have a more complete understanding of the research problems / questions" such as "comparing different perspectives drawn from quantitative and qualitative data" or "understanding experimental results by incorporating the perspectives of individuals".

Typically, the starting part for linguistic analyses is a text-based corpus. A linguistic corpus is "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2005). A corpus is made for the study of language phenomena. The corpus may already exist or it may need to be constructed, depending on the research question.

Sinclair (2005) provides an indepth guide to good practice when it comes to developing linguistic corpora in Wynne's 'Developing Linguistic Corpora: a Guide to Good Practice' (2005)[1].

Step 3: Extract (domain) understanding

When domain understanding is referred to in the context of this work, it means expertise or what Lundvall and Johnson (1994) refer to as 'know-how'. Jarvenpaa and Staples (2001) define expertise as "involving intangible information embodied in human memory, knowledge, experience or skill". They contrast this with the concept of information which they define as "an information product such as a written document or computer program".

[1]How to build a corpus: `https://ota.ox.ac.uk/documents/creating/dlc/appendix.htm`

This part of the process aims to capture humanistic insight from a dataset. DH research often involves taking a dataset and carrying out interpretive work in order to draw meaningful insight from it.

In the case of the *Industrial Memories* project, "the narrative form of the Ryan Report is deconstructed and key information extracted" such as every individual named in the report using Stanford Named Entity Recognizer (NER) system[2]. This information was then used to create a collocation network of entities within the Ryan Report (Leavy, Pine, and Keane, 2018).

Step 4: Knowledge codification

Knowledge codification involves representing knowledge in a way that it can be reused. Humanities data is characteristic in the fact that it is generally not as mathematically measurable as scientific data (Di Cresce and King, 2017). Therefore, in this process, it is imperative to codify the data or understanding extracted in step 3 of the process in order for it to be reused in a meaningful way by the receiving discipline (D.2).

There are various ways of codifying knowledge extracted from humanities data sources. Corpus annotation is the practice of adding interpretative linguistic information to a corpus (Leech, 2005). Depending on the type of linguistic exploration, different types of annotation may be appropriate.

One of the main ways in which knowledge or understanding is codified in Digital Humanities is through XML or TEI encoding. The Text Encoding Initiative (TEI) is a consortium which maintains and develops the TEI *Guidelines for Electronic Text Encoding and Interchange* which "define and document a markup language for representing the structural, renditional, and conceptual features of texts"("TEI: Guidelines"). The XML based TEI markup language has been widely adopted in the Digital Humanities community, particularly when creating digital editions such as Transcribe Bentham[3] which was presented in Chapter 2.

[2]Stanford Named Entity Recognizer: `https://nlp.stanford.edu/software/CRF-NER.shtml`

[3]Transcribe Bentham: `http://blogs.ucl.ac.uk/transcribe-bentham/`

The guidelines are designed to be customised for specific projects and there is extensive documentation of how to produce TEI customisations as well as sample methods and customisations which can be consulted. The guidelines comprise an XML schema which is available in many formats (DTD, Relax NG and W3C XML schema) and Piotrowski (2012) believes it is "probably one of the largest XML schemata in existence. There is a recognition that the disciplinary needs of the TEI community are diverse and customisation of the language is important to ensure its expressiveness. Rather than prescribing exact constraints for its users, it provides a general framework which can be customised by choosing the elements required. There is a core TEI module which contains the most common elements of texts. The core module can be augmented with additional modules depending on the text type, for example correspondence, poetry, drama or transcribed speech (Piotrowski, 2012).

Step 5: Collaborate to evaluate

Collaboration is essential in interdisciplinary research given the complexity of real-world problems (Moreno, Kynčlová, and Werthner, 2016). Evaluation occurs where effectiveness is examined or judged systematically and empirically (Patton, 1990). In the case of the relo-KT process, it is necessary to evaluate the quality of the knowledge codification. The aim of codifying the knowledge in a systematic way is so that it may be reused - potentially by a researcher(s) from the same or another discipline. Rockwell (2012) notes that it "is common in electronic text projects to bring in consultants to review encoding schemes and technical infrastructure — such expert consultations should be budgeted into projects in order to make sure projects get outside help, but they can also serve as formal, though formative, opinions on the excellence of the work". DH and digital scholarship is a relatively new discipline, and as a result, there is "an absence of peer review mechanisms" for many types of digital work (Rockwell, 2012).

Flanders (2013) articulates the question of how to evaluate digital scholarship as follows:

> "digital scholarship reveals a conundrum that has lain at the heart of humanities scholarship for decades: how can we simultaneously encourage paradigm shifts and radical revisions of our

> modes of analysis, and also know how to evaluate them once we have them before us"

Much of the literature on evaluation in DH centres around the evaluation of digital scholarship in terms of peer review, particularly for tenure and promotion (Anderson and McPherson, 2011; Cavanagh, 2012; Mattern, 2012; Nowviskie, 2012; Rockwell, 2012; Schreibman, Mandell, and Olsen, 2011) and how to "translate existing guidelines like those of the Modern Language Association MLA into terms and practices that are meaningful for our home departments and universities".

Huutoniemi (2010) raises questions about evaluating interdisciplinary and transdisciplinary research such as:

- How can balance be achieved between different epistemic viewpoints and what criteria may be used to assess them?
- How should the evaluation of research be organised?
- Who should be included in the evaluation process?

Huutoniemi does not provide a framework for evaluating cross-disciplinary research, but rather presents some of the key values which relate to interdisciplinary research. These are used to suggest different approaches for assessing the quality of interdisciplinary research. Essential elements when it comes to evaluating cross-disciplinary research include "excellence" within the individual disciplines and an iterative cycle communicating tacit knowledge.

In the projects discussed in Chapter 2, the types of evaluation carried out varied depending on the project. The *CULTURA* project merits closer discussion in terms of its evaluation approach in this section, as rather than focusing solely on the scholarly output, the system was also evaluated. As detailed by Steiner et al. (2014), the system was evaluated using a multi-method approach which was developed in accordance with state of the art. Methods including questionnaires, focus group discussions and log data allowed quality assessment of the *CULTURA* virtual research environment. Both the

Old Weather and *Transcribe Bentham* are evaluated by human-assessed quality control.

A key part of the relo-KT evaluation process is the interaction between the domains (D.1 and D.2). As previously outlined, it is recognised that this is a cyclical, iterative process in which D.2 can implement change in the format and content of the codified knowledge.

Step 6: Repository

In computing, a repository is "a central location in which data is stored and managed" (Oxford Dictionary). Public repositories are visible to everyone which means that the data stored within can be accessed and reused by others. In his 'Short Guide To Evaluation Of Digital Work', Rockwell (2012) advises that unless digital work is "documented and deposited" in a "well-managed" repository, there is a risk that a generation of digital scholarship may be lost. Rockwell (2012) also notes that well managed repositories were "just emerging" in 2012. In the intervening period, there has been a lot of development in data repositories and there are now a wide variety of repositories to deposit research-related data. These include institutional repositories, national digital repositories, research repositories and software repositories.

3.4 Summary

The process described in Chapter 3 is a combination of mixed methods for cross-disciplinary knowledge transfer. In Chapter 5, relo-KT will be applied to a particular use case to transfer domain understanding of rhetorical devices to MT practitioners. In the RF-MT use case (Rhetorical Figure-MT), D.1, refers to linguistics. The linguist holds the domain understanding and engages in a collaborative process of knowledge transfer and negotiation with MT researchers (D.2).

Table 3.1 outlines a depiction of the process as it was presented in this chapter.

Step	Possible approaches
1. Establish question	collaborate with experts from D.2
2. Acquire data	assemble a corpus - textual - visual - audio extract data from data sources create a model - 3D
3. Extract domain understanding	text mining manual analysis spatial analysis visual analysis
4. Codify knowledge	annotate data - XML markup - TEI-XML markup - image tagging spreadsheets database
5. Collaborate to evaluate	run a test / algorithm ask experts check by hand
6. Build repository	follow disciplinary guidelines for research data management

TABLE 3.1: Description of the relo-KT process

Chapter 4

Rhetorical Figures and Machine Translation

> A major difficulty in translation is that a word in one language seldom has a precise equivalent in another one
>
> *Arthur Schopenhauer*

4.1 Introduction

In Chapter 2, literature related to interdisciplinary research and knowledge transfer (KT) was presented and discussed. From that literature review, the relo-KT process for cross-disciplinary KT was developed in Chapter 3. In Chapter 5, relo-KT will be put to the test in a use case which takes the linguistic understanding of a sample of persuasive rhetorical figures from a corpus of political speech and transfers it to the MT domain (the RF-MT use case). To contextualise the RF-MT use case, the literature pertaining to rhetorical figures and persuasive language (D.1 - linguistics) and how they relate to the MT domain (D.2) is reviewed.

4.2 Domain 1: linguistics

Similar to the notable examples of Martin Luther King's "I have a dream" and Winston Churchill's "we shall fight" speeches which were mentioned in

Chapter 1, there are numerous other examples of persuasive political language which have permeated our cultural consciousness[1]. When Barack Obama stood on stage in Nashua, the night after the New Hampshire primary in the 2008 US presidential campaign, and uttered the words "Yes we can", he was not simply uttering a campaign slogan. He was attempting to persuade the American electorate that *he* was the candidate who could change the style of politics that they had grown tired of during the leadership of his predecessors. One of the ways he achieved this was by employing the art of persuasion, in the form of rhetorical figures, to connect emotionally with voters. In the speech, often since referred to as the "Yes we can" speech, Obama uttered the mantra eleven times and "Yes we can" went on to become the defining phrase of Obama's successful presidential campaign. Whilst not all political rhetoric is as effective and memorable as this example, politicians and political speech writers tend to employ rhetorical figures when preparing statements and speeches. They tend to be interwoven seamlessly in the oration, yet carry an important function: to persuade. For this reason, understanding these linguistic devices and presenting that understanding in a way that is consumable for those who work with rapidly advancing technology (such as MT) is of value. This section presents a comprehensive background of how rhetorical figures are used for persuasion.

4.2.1 The language of persuasion

> Advertising is fundamentally persuasion and persuasion happens to be not a science, but an art. Advertising is the art of persuasion.
>
> *William (Bill) Bernbach*
> *American Advertising Executive, 1960s/1970s*

For millennia, people have studied and paid attention to the art of rhetoric. Since Ancient Greece, public speakers have practised the art of rhetoric and it is an art which has endured and can be observed in the speeches of the most

[1]Cultural consciousness can be defined as the process of developing awareness of culture in the self, which can result in expanding understandings of culture and developing deeper cultural knowledge about other individuals and contexts (Páez and Albert, 2012)

recent American presidential election in 2016 (Martin, 2016) and Britain's 2016 'Brexit' referendum campaign (Crines, 2016).

Rhetorical figures tend to fall under one of two classifications - schemes or tropes. Charteris-Black (2011) distinguishes between schemes and tropes by describing a trope as "a figure of speech in which words are used in a sense different from their literal and normal meaning" while schemes tend to "concern the arrangement or sequencing of words that affect a sentence's structure". In narrowing down the focus of this study, the decision was made to focus on schemes (order and syntax) rather than tropes (the meaning of words).

Despite the millennia spent observing the phenomenon of rhetoric, to date rhetorical figures have not been widely considered in machine learning processes. There is a limited bank of exemplar rhetorical figures and the ones which are most often cited tend to come from famous speeches, poetry and works of literature. Harris et al. (2018) reference Fahnestock (1999) to make the following point in relation to the correlation between linguistic forms and rhetorical functions:

> "it is still an open question how well the form–function couplings that humanists have found stand up beyond the small sampling of discourse they have cared to explore – the orations, poetry, and killer ripostes that stock their examples."

That said, the Silva Rhetoricae resource has identified in excess of 400 figures in the English language, many of which can be used for the purpose of persuasion[2]. Some of these are explored in the next section.

[2]Silva Rhetoricae: http://rhetoric.byu.edu/

4.2.2 Rhetorical figures in political speech

Persuasion is a fundamental strategy in politics and skilled speech writers employ a range of rhetorical figures. Often, these figures are used to indirectly influence an audience to accept a politician's argument (David, 2014). They can be used for a variety of purposes including (but not limited to):

- **Inspirational** purposes - Winston Churchill's inspirational "we will fight on the beaches" speech led to an "extraordinary 88% approval rating" despite the incessant bombardment of London that was underway during World War 2 (Bungay, 2009)
- To **encourage** - the use of the "Yes We Can" refrain in Obama's 2008 presidential election campaign encouraged people from different cultural, religious, racial and gender backgrounds to participate in civic engagement (Bang, 2009) which not only saw a forty-year high in voter turnout, but was also "one of the most diverse in US history" (Lopez and Taylor, 2009)
- To **influence** - Hilary Clinton's 1995 "Women's rights are human rights" speech still resonates two decades later. Despite the fact that "women and girls had made progress in health and education", they still lag "in political rights and security" (Chozick, 2015)

Rhetorical figure examples

As previously mentioned, there is an abundance of rhetorical figures which can be used for persuasion. However, not all are commonly used. Epanaphora, epistrophe, polyptoton and polysyndeton were the four figures chosen for examination in this research as they commonly occur in political speech, they are detectable by Rhetorica and findings from a pilot study demonstrated that they are present in the context of Irish parliamentary speech.

Epanaphora

Burton (*Silva Rhetoricae: The Forest of Rhetoric*) defines epanaphora (sometimes referred to as anaphora) as the "repetition of the same word or group of words at the beginning of successive clauses, sentences, or lines". Epanaphora is often used for emphasis. A message that Hilary Clinton wanted to "echo forth" from her speech at the 1995 United Nations Fourth World Conference on Women in Beijing was that "Women's rights are human rights"[3]. To aid the delivery of her message that "as long as girls and women are valued less" peace and prosperity in the world would not be attainable, she used epanaphora to list ways in which women's human rights can be violated:

> **It is a violation of human rights when** babies are denied food...
>
> **It is a violation of human rights when** women and girls are sold...
>
> **It is a violation of human rights when** women are doused...
>
> **It is a violation of human rights when** individual women are raped...
>
> **It is a violation of human rights when** a leading cause of death...
>
> **It is a violation of human rights when** young girls...
>
> **It is a violation of human rights when** women are denied the right...

Clinton's repetition of the first eight words of each sentence has the effect of creating a relentless rhythm in the text which ultimately makes the phrase more memorable for the listener. It also gives a concise structure to the seven

[3]Hilary Rodham Clinton, Remarks to the United Nations 4th World Conference on Women: https://www.americanrhetoric.com/speeches/hillaryclintonbeijingspeech.htm

violations that she mentions. The repetition, along with the gruesome imagery she evokes, prompts the audience to reflect on the unjust and unfair reality that many women around the world are born into (Johnson, 2015).

Epistrophe

The Charleston church shooting was a racially motivated mass shooting in which nine African American people were murdered during a prayer service in Charleston, South Carolina in June 2015. Nine days after the shooting, President Barack Obama delivered a eulogy at the Emanuel African Methodist Episcopal Church, at which it had taken place. The eulogy he delivered on 26 June 2015 was described in The New York Times as "remarkable" because it "drew on all of Mr. Obama's gifts of language and empathy and searching intellect" (Kakutani, 2015).

Throughout the address, Obama wove the theme of grace skilfully and eloquently. Towards the end of the eulogy, he began to sing lines from the hymn *Amazing Grace*:

> Amazing grace,
> how sweet the sound,
> that saved a wretch like me;
> I once was lost,
> but now I'm found;
> was blind but now I see.

and followed this directly by naming each victim of the shooting and describing how they had "found that grace".

> Clementa Pinckney **found that grace**.
>
> Cynthia Hurd **found that grace**.
>
> Susie Jackson **found that grace**.
>
> Ethel Lance **found that grace**.
>
> DePayne Middleton-Doctor **found that grace**.

> Tywanza Sanders **found that grace**.
>
> Daniel L Simmons, Sr. **found that grace**.
>
> Sharonda Coleman-Singleton **found that grace**.
>
> Myra Thompson **found that grace**.

Epistrophe is defined by Burton (*Silva Rhetoricae: The Forest of Rhetoric*) as "ending a series of lines, phrases, clauses, or sentences with the same word or words". Obama's use of epistrophe in the Charleston eulogy is simple, yet extremely effective. By repeating the word grace from the song through the speech, it enabled him "to pay a powerful tribute to the individual victims of the tragedy" by commemorating them in the musical tradition of their religious community (*LitCharts*).

Polyptoton

A common rhetorical figure used to persuade in speech, political or otherwise, is polyptoton, in which a word is repeated in close proximity, but in a different form[4]. A celebrated example of polyptoton is from US President John F. Kennedy's inaugural speech delivered on 20 January 1961:

> Now the trumpet summons us again — not as a call to bear arms, though arms we need; not as a call to **battle**, though **embattled** we are — but a call to bear the burden of a long twilight struggle, year in and year out, "rejoicing in hope, patient in tribulation" — a struggle against the common enemies of man: tyranny, poverty, disease, and war itself.

With this use of polyptoton, Kennedy acknowledges the ongoing cold war between the US and the USSR while at the same time "calling on American citizens to act for reasons beyond that battle" (*LitCharts*).

[4]Repeating a word, but in a different form. Using a cognate of a given word in close proximity (*Silva Rhetoricae: The Forest of Rhetoric*)

Polysyndeton

On 28 January 1986, the American space shuttle Challenger exploded and broke apart shortly after lift off, resulting in the deaths of its seven crew members. This catapulted the American nation into a state of mourning as many citizens witnessed the accident live on television. President Ronald Reagan postponed his annual State of the Union address scheduled for the same day, and instead delivered what has become one of the most significant American speeches of the 20th century (ranking 8th on the top 100 speeches on the American Rhetoric website[5]).

Polysyndeton is defined as "employing many conjunctions between clauses, often slowing the tempo or rhythm" (*Silva Rhetoricae: The Forest of Rhetoric*). The following example from Reagan's address to the nation on the explosion of the Space Shuttle Challenger features the repetition of **and**:

> We will always remember them, these skilled professionals, scientists **and** adventurers, these artists **and** teachers **and** family men **and** women, and we will cherish each of their stories—stories of triumph and bravery, stories of true American heroes.

Polysyndeton was used in this case "to lend gravitas" to the speech. The natural pauses which polysyndeton inserts into the flow of speech gave the listeners a chance to envisage the victims "as a diverse group of individuals, who had families, professions, and goals" and enabled him to convey "the human scale of the tragedy" (*LitCharts*).

The four quotations highlighted above are oft-cited examples of rhetorical figures in action in political speech. Mostly delivered by American presidents, they would have been crafted and honed to reinforce a particular message.

[5] American Rhetoric Top 100 Speeches:
`https://www.americanrhetoric.com/top100speechesall.html`

4.2.3 Computational rhetoric

Computational rhetoric is a nascent interdisciplinary field which brings together linguistics and rhetoric with computer science and natural language technology. As previously mentioned, rhetoric has a long history which has "concerned itself with pragmatic function, style, and affect in language" (*Computational Rhetoric Workshop*). These are topics which are key components of the computational linguistics field, and from this field have come areas of investigation such as 1, 2 and argument(ation) mining (Lawrence, Visser, and Reed, 2017).

As Peldszus and Stede (2013) note "one of the central aspects of human communication is argumentation: the process of conveying inclinations, attitudes or opinions" and attempting to make another party accept - or even adopt - them. Argument(ation) mining pursues "the automatic identification of the argumentative structure contained within pieces of natural language text" (Lawrence, Visser, and Reed, 2017). By identifying this argumentative structure, Lawrence, Visser, and Reed claim that they "are able to tell not just what views are expressed, but also why they are held".

The following sections explore key aspects of computational rhetoric in more detail.

4.2.4 Automatic detection and annotation of rhetorical figures

Despite the advances in machine learning technology, there exists a dearth of annotated rhetorical figure data which can be exploited for ML purposes. In an attempt to alleviate this shortage, Harris et al. (2018) have developed a markup scheme which they hope will lead to the resolution of the "annotated data bottleneck" identified by Dubremetz (2017):

> "if we want to tap into the resources of ML to meet the challenge of Rhetorical Figure combinatorics, we will need texts annotated

> for occurrences of multiple figures – mutually re-enforcing, often interpenetrating, sometimes wholly overlapping multiple Rhetorical Figures, but also independent, or even mutually interfering, multiple Rhetorical Figures as well. That prospect requires a standardized annotation scheme. We have developed such a scheme." (Harris et al., 2018)

Before focusing closely on Harris et al. (2018)'s Waterloo Annotation Scheme for Rhetorical Figures (WAS), it is important to present some of the fundamental research which has been carried out on the automatic detection and annotation of rhetorical figures (Dubremetz, 2017; Gawryjolek, 2009; Harris et al., 2018; Hromada, 2011; Java, 2015; O'Reilly and Paurobally, 2010). A review of the literature related to this topic reveals that while there is a focus on machine learning (ML), there is little published work which takes place at the intersection of rhetorical figure detection, annotation and MT. This is a gap that will be explored more in the subsequent chapters of this thesis.

Gawryjolek (2009) described the annotation of rhetorical figures as a "new problem of linguistic annotation" and developed JANTOR (Java ANnotation Tool Of Rhetoric) which combined Stanford's parser API[6] with two APIs for WordNet searching: a Java API for WordNet Searching (JAWS)[7] and the MIT Java WordNet Interface (JWI)[8] to both annotate and navigate rhetorical figures within a corpus.

Hromada (2011)'s work which proposed a method for extracting figures of speech was based upon a translation of a canonical form of repetition-based figures of speech into the language of PERL-compatible regular expressions. Hromada (2011) dealt with four rhetorical figures which matched more than 7000 strings when applied on dramatic and poetic corpora written in English, French, German and Latin. Hromada's work expands on Gawryjolek's which is "operational only when combined with [a] probabilistic context-free grammar parser adapted to [the] English language", therefore does not function when applied on languages for which the parser does not exist.

[6]Stanford JavaNLP API Documentation: https://nlp.stanford.edu/nlp/javadoc/javanlp/
[7]Java API for WordNet Searching (JAWS): https://github.com/jaytaylor/jaws
[8]JWI: https://projects.csail.mit.edu/jwi/

Java (2015) picked up on Gawryjolek (2009)'s work and developed it to create a software tool called Rhetorica which identifies 14 classical rhetorical figures in free English text. Java used measures of classical rhetorical structure to improve accuracy for authorship attribution tasks. In Chapter 5, Rhetorica will be presented and discussed at length with regard to the RF-MT use case.

Much of the literature around rhetorical figure annotation features work by Randy Harris[9] and Chrysanne DiMarco[10] from the University of Waterloo in Canada, including the construction of a rhetorical figure ontology (Harris and DiMarco, 2009). Harris et al. (2018) developed an XML annotation scheme for rhetorical figures in an attempt to alleviate the problem caused by the "lack of annotated data" highlighted by Dubremetz (2017). The Waterloo Annotation Scheme for Rhetorical Figures (WAS) is an XML based scheme which allows for straightforward, scalable and adaptable markup of rhetorical figures (Harris et al., 2018). The scheme identifies the rhetorical figure itself and its defining elements. It uses the name of the figure to open the XML markup and for each element, it uses the figure name followed by "an alphabetic variable, to distinguish the elements; followed by a numeric designation, to mark the sequence in which the elements occur; separated by dashes; all terms mandatory" (Harris et al., 2018).

The WAS markup in action is demonstrated on Churchill's 'We Shall Fight' epanaphora:

```
<epanaphora>
  <epanaphora-A-1> we shall fight </epanaphora-A-1> in France,
  <epanaphora-A-2> we shall fight </epanaphora-A-2> on the seas and oceans,
  <epanaphora-A-3> we shall fight </epanaphora-A-3> with growing confidence
      and growing strength in the air,
  we shall defend our Island, whatever the cost may be,
  <epanaphora-A-4> we shall fight </epanaphora-A-4> on the beaches,
  <epanaphora-A-5> we shall fight </epanaphora-A-5> on the landing grounds,
  <epanaphora-A-6> we shall fight </epanaphora-A-6> in the fields and in the
      streets,
  <epanaphora-A-7> we shall fight </epanaphora-A-7> in the hills; we shall
      never surrender...
```

[9]Randy Harris:
https://uwaterloo.ca/english/people-profiles/randy-harris
[10]Chrysanne DiMarco:
https://uwaterloo.ca/english/people-profiles/chrysanne-di-marco

```
10 </epanaphora>
```

Annotating rhetorical figures is a challenging task. These linguistic devices often overlap or one figure can be read as another depending on the context. "Overlap is the common term for cases where some markup structures do not nest neatly into others" (DeRose, 2004). An example of overlap exists in Barack Obama's New Hampshire Primary Concession speech, often referred to as his '*Yes we can*' speech[11]. To make his argument, Obama weaves together epanaphora (the repetition of the same word or group of words at the **beginning** of successive clauses, sentences, or lines) and its mirroring rhetorical figure, epistrophe (the repetition of the same word or group of words at the **end** of successive clauses, sentences, or lines):

> **It was** a creed written into the founding documents that declared the destiny of a nation: ***Yes, we can***.
>
> **It was** whispered by slaves and abolitionists as they blazed a trail towards freedom through the darkest of nights: ***Yes, we can***.
>
> **It was** sung by immigrants as they struck out from distant shores and pioneers who pushed westward against an unforgiving wilderness: ***Yes, we can***.
>
> **It was** the call of workers who organized, women who reached for the ballot, a President who chose the moon as our new frontier, and a king who took us to the mountaintop and pointed the way to the promised land: Yes, we can, to justice and equality.

Additionally, the combination of epanaphora and epistrophe is a rhetorical figure in its own right: symploce involves "beginning a series of lines, clauses, or sentences with the same word or phrase while simultaneously repeating a different word or phrase at the end of each element in this series" (*Silva Rhetoricae: The Forest of Rhetoric*).

The nature of these figures means that there is often a requirement to markup several, sometimes overlapping, figures within a text. Gawryjolek (2009) and Harris et al. (2018)'s solution to this problem is to use a system of standoff

[11]The speech can be read in full on: `https://www.americanrhetoric.com`

markup. Rather than tagging every rhetorical figure inline within the text, all information concerning annotations is saved in a separate XML file to the text. This enables markup of multiple rhetorical figures within the same document, while the original text remains unchanged. Using standoff markup allows for scalability and adaptability, ultimately allowing for expansion.

4.3 Domain 2: machine translation

Machine translation (MT) is one of the oldest and most challenging problems facing AI. As Doug Arnold (2003) observed, MT is difficult because translation itself is a complicated task. A translator must take a text in one language (the source) and produce an equivalent text in another language (the target language). Often, the expectation on a translator is not simply to produce a text which is equivalent in meaning, but there is an expectation that the target text will also have equivalence in style. It is understandable that a translation of a literary work is expected to be as "clear, unambiguous, interesting, persuasive, elegant, poetic, gripping, etc." as possible (Arnold, 2003).

MT has come a long way since its inception and while improving it "remains a challenging goal" (Quoc and Schuster, 2016), any advances made have the potential to make a large impact and have genuine implications for the field. In this section, I will first provide a brief synopsis of the history of MT, before giving an overview of the MT landscape in 2018.

4.3.1 History of MT

Early MT systems followed a rule-based approach which combined large bilingual dictionaries and rules related to the morphology and syntax of the source and target languages. However, the rule-based approach has a number of drawbacks. It is quite restrictive and has been in decline since MT research has married with other technologies, creating more intelligent hybrid approaches such as statistical MT (SMT) and neural MT (NMT).

Rule based MT (RBMT)

Three main systems come under the RBMT umbrella.

Direct MT systems translated text from source to target language word by word with very little analysis of the source text. These systems relied heavily on large bilingual dictionaries. For each word in the source language, the dictionary specifies a set of rules for translating it. Once the words have been translated, simple rules are applied to reorder certain words such as placing adjectives after nouns when translating from English to Spanish. However, problems would arise when it came to word order in long sentences or words were translated without any knowledge of their syntactic role in the sentence.

Transfer based MT - the problems which occured with direct MT systems led to a development of transfer based MT which worked on a sentence or phrase level as opposed to the word level. Transfer based MT uses a three-step approach to translation. First the system would analyse the source language sentence using a semantic or syntactic representation of that langauge (eg. a parse tree). Stage 2 was the transfer stage in which the source language representation would be transformed into the matching target language representation using a set of rules. The translated sentence in the target language is then generated from the language representation in stage two.

The **interlingua** approach "eliminates the transfer component" by taking a language-independent approach (Arnold, 2003). However, this approach still requires an analysis stage which relies on complex rules, and "one cannot expect analysis and synthesis rules for one language to be identical" (Arnold, 2003). It is very difficult to create a language independent representation given that each language is idiosyncratic and no two languages perceive concepts in exactly the same way.

In their survey of current paradigms in machine translation, Dorr, Jordan, and Benoit (1999) stated that RBMT "systems fail for texts that rely heavily on metaphor and world knowledge because they have great difficulty in representing and using complex and subtle metaphors or understanding social context and interactions, and it is nearly impossible for them to keep up with the rapid changes in vocabulary".

Example based MT (EBMT)

The next major step in the MT timeline is example-based MT (EBMT). EBMT emerged in the 1990s as an alternative to RBMT systems (Somers, 2001). Rather than following a pre-determined set of rules, EBMT uses a bilingual corpus as the source of linguistic knowledge "examples" of previous translations. New translations are made by searching the corpus to find the most similar example from what has already been translated. It then uses this as a model for the new translation using a process known as "recombination" (Somers, 2001; Turcato and Popowich, 2003).

Statistical MT (SMT)

EBMT paved the way for SMT which relies on parallel corpora to train the translation examples. "Statistical MT employs two distinct and separate processes: training and decoding" (Hearne and Way, 2011). The machine learns a certain amount automatically by examining large amounts of parallel text documents which are nearly exact translations of each other. The Canadian government's parliamentary proceedings (Hansard) is one such resource as it is bilingual (English and French). It is drawn from official records and "it spans a broad assortment of topics and the stylistic range includes spontaneous discussion and written correspondence along with legislative propositions and prepared speeches" (Roukos, Graff, and Melamed, 1995). Another example is the parallel EuroParl corpus from the proceedings of the European Parliament which initially covered 11 language pairs (Koehn, 2005), but the latest version (2012) has increased to include 21 European languages. One of the challenges of SMT is that parallel corpora do not exist for many language pairs.

Neural MT (NMT)

The most promising breakthrough to date in the MT field has arguably been neural MT (NMT). In a shift away from statistical methods, neural

networks are employed to address the shortcomings of more traditional MT approaches. Neural models "involve building an end-to-end neural network that maps aligned bilingual texts" (Castilho et al., 2017) and uses sequence prediction to produce a target sentence (*The State of Neural Machine Translation (NMT)*).

However, NMT still hasn't proven to be the panacea to MT's shortcomings and NMT systems tended to fare worse in accuracy than traditional systems. Wu et al. (2016) attribute this to three inherent weaknesses of NMT:

1. *slower training and inference speed*: it takes a considerable amount of time and computational resources to train a NMT system on a large-scale translation dataset which slows the rate at which experimentation and innovation can be achieved
2. *ineffective when rare words are encountered*
3. *failure to translate all words in the source sentence*: NMT systems sometimes "fail to completely "cover" the input, which can result in surprising translations" (Wu et al., 2016)

To address the deficiencies outlined above, Google announced their GNMT system in 2016 which achieves "the largest improvements to date for machine translation" using state of the art training techniques (Wu et al., 2016). On its launch, the developers of GNMT claimed that their system delivered approximately 60% fewer "translation errors on several popular language pairs".

To sum up, despite the many advances in the field of MT, shortcomings remain and even the most advanced MT systems can still make significant errors like "translating sentences in isolation rather than considering the context of the paragraph or page" (Quoc and Schuster, 2016). Fernand Cohen (2018) asserts that "the biggest challenge for human translators and even more so for MT is the ability to correctly understand the cultural nuances of what is written or said" in a source text.

4.4 Discourse MT

Discourse analysis is the study of language beyond the sentence level. It is "the study of real language use, by real speakers in real situations" (Van Dijk, 1985). The interconnection between MT and discourse analysis is one which has been developing since the first workshop on Discourse in Machine Translation (DiscoMT) was held in 2013 (*Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*). The workshops mainly focus on language processing techniques, but for 2019, there is special emphasis "on how well neural MT systems handle discourse-level phenomena". Discourse MT involves taking discourse phenomena such as dropped pronouns into consideration[12]. In 'On Integrating Discourse in Machine Translation', Karin Sim Smith (2017) provides a comprehensive outline of the recent work that has been undertaken in the area of Discourse MT. Smith highlights that despite improving the quality of discourse in automatic translations, and an increase in research being carried out on discourse in MT, "many of the existing models and metrics have yet to integrate these insights".

Smith claims that "in order to take MT to another level, it will need to judge output not based on a single reference translation, but based on notions of fluency and of adequacy – ideally with reference to the source text". Moving beyond MT at the sentence level is a challenge which MT researchers such as Hardmeier (2012) have attempted to address and found that "even advanced MT systems still assume that texts can be translated sentence by sentence and that the sentences in a text are strictly independent of one another". Libovicky and Cartoni (2018) argue that in order "to translate entire paragraphs and documents consistently, i.e. in a lexically coherent and pragmatically appropriate manner. Argumentative structure of text, consistency of lexical choice, and the right 'tone' for its pragmatic intent are the next problems to focus on".

Smith (2017) acknowledges that "one of the problems with repetition is indeed automatically recognising where it results in consistency, and where it works to the detriment of lexical variation". Garcia et al. (2017) attempt

[12]Dropped pronouns: in which pronouns are frequently dropped in the source language but should be retained in the target language (Wang et al., 2016)

to improve the quality of translation by taking into account that certain linguistic phenomena such as discourse markers go beyond the boundary of a sentence. Consistency of such devices is "difficult to attain if the document is translated in a sentence by sentence basis". To do this, they use a cosine similarity metric between word embeddings to check if they are semantically similar, in an attempt to encourage consistency for the same word to be translated in a similar manner throughout the document. They found that "although differences among systems are not statistically significant for the automatic evaluation metrics, they are noticeable for human evaluators that prefer the outputs from the enhanced systems".

4.5 MT evaluation

Attaining quality MT evaluation is one of the most challenging tasks for the field. There are two main evaluation approaches in MT. Human evaluation is reliable, but costly. It is usually only used for the final evaluation to verify that the final translation reaches an adequate standard. Automated evaluation on the other hand, makes a more cost effective solution which tends to be used in development, as the results are not as robust as with human evaluation. BLEU (bilingual evaluation understudy) is a metric which is used for evaluating the quality of text which has been machine-translated from one natural language to another (Papineni et al., 2002). BLEU is widely used in MT evaluations, although there is an argument that the MT community is over-reliant on it as a metric (Callison-Burch, Osborne, and Koehn, 2006).

Additional difficulties persist when it comes to evaluating the quality of automatically translated discourse phenomena and concludes that although "the translator's role as mediator will not easily be replaced by machines... we must ensure we assess MT output based on a measure of adequacy compared to the *source*, if it is to fulfil its purpose in terms of communication" (Smith, 2017).

The solution to these difficulties is often to put a 'human in the loop' of the machine processes.

4.6 Post-editing

Language translation carried out by humans is a slow and expensive process. Green, Heer, and Manning (2013) calculated that to translate the entire CHI (ACM Conference on Human Factors in Computing Systems) Proceedings from 1982 to 2011 from English to just one other language would cost approximately $2.2 million. For this reason, in recent years, there has been an increasing demand for post-editing MT output and this has "led to a propagation of research" (O'Brien and Simard, 2014).

Post-editing involves the manual correction of texts which have been translated from a source language into a target language by an MT system (Allen, 2001). Typically, this involves "tidying up the raw output, correcting mistakes, revising entire, or, in the worst case, re-translating entire sections" (Somers, 2003). The most common approach to post-editing involves using a tools as the editing environment. PET is a stand-alone, open-source tool to post-edit and assess machine or human translations while gathering detailed statistics about post-editing time and other effort indicators. The tool is a graphical user interface which displays the source and the translation texts side by side to facilitate post-editing (Aziz, Castilho, and Specia, 2012).

4.7 MT, transcreation and persuasion

Rhetorical figure translation is an under-explored area in both human translation (HT) and MT but some research has been carried out, primarily in the transcreation[13] domain. Traditionally, translators attempt to stay as faithful to the source text as possible. Transcreation differs in that the intercultural context is taken into consideration when translating a text to a target language (Katan, 2016). Pedersen (2014) describes transcreation as "a translation-like activity" which is predominantly used to refer to "the adaptation of advertising material for different markets" while Merino (2006) and

[13]The process of adapting a message from one language to another, while maintaining its intent, style, tone and context https://en.wikipedia.org/wiki/Transcreation

Rike (2013) both focus on the creative aspect of trans**CREAT**ion. To quote Shriver (2011):

> In transcreation, translators aim to produce a conversion that stays close [to the original], while also evoking the desired reaction from those who receive the message in the target language. Transcreation involves neither a strict translation nor creation of a message from scratch.

A number of studies have examined the role of transcreation in developing healthcare educational materials for Hispanic populations in the US (Solomon et al., 2005), (Simmons et al., 2011), (Wells et al., 2013), (Rivera et al., 2016). Take for example Rivera et al. (2016)'s transcreation of cancer educational material in the US territory of Puerto Rico, where it is claimed that "cancer is the leading cause of death" (Rodríguez, 2012). Rivera et al. (2016) used a collaborative approach to review the Spanish version of the *Cancer 101* curriculum[14] to determine whether it could be used to provide cancer education in Puerto Rican communities. In addition to making their anticipated changes to words, images and statistics (such as cancer statistics for the general US Hispanic population), they received suggestions to further tailor the materials for the communities that would be using them.

While not directly related to the translation of rhetorical figures and political speech, transcreation demonstrates the levels of nuance that have to be taken into consideration when translating. An eloquent translation of the words and sense of the text may not be sufficient in a life or death scenario such as healthcare for vulnerable patients.

[14]The *Cancer 101* curriculum is a cancer education resource developed in collaboration with American Indians/Alaska Natives to improve cancer knowledge, action regarding cancer control in tribal settings, and survival rates for members of their communities (Hill et al., 2010)

4.8 MT and rhetorical figures

Some of the work carried out to date on the translation of rhetorical figures has taken place in the context of advertising, a domain which relies heavily on persuasion to sell products, concepts or brands. Repetition has been used in advertising as a method to increase brand familiarity and make a product stand out over its competitors (Campbell and Keller, 2003). Repetition comes in various forms. For example, images of a product or brand logo can be placed in a number of locations within the same advertisement. A slogan may be repeated across subsequent advertising campaigns so that it becomes associated with a product or brand (e.g. McDonald's "I'm loving it" and Nike's "Just do it") or the words and phrases within the advertisement repeat in some way such as Chevrolet's "Eye it. Try it. Buy it." campaign which uses the repetition of 'it' to have a rhyming effect and also uses a parallelism which in this case is a tricolon: "three parallel elements of the same length occurring together in a series" (*Silva Rhetoricae: The Forest of Rhetoric*).

In advertising, the headline plays a key role in persuading the reader to firstly continue to read the advertisement and ultimately go on to buy the product. In a study on how rhetorical figure usage is dealt with in English to Russian translations of magazine advertisement headlines, Smith (2006) outlines that:

> "Translation theory suggests that advertising texts should be translated to create a target-language advertisement which will have a positive impact on the target audience. It is thus not of primary importance whether a particular rhetorical figure is translated by the exact same figure in Russian; what is important is that the target-text headline should have the same attention grabbing function as the original"

To demonstrate this, figure 4.1 provides an example of an English advertising slogan for L'Oréal Paris' Color Riche™ lipstick which was advertised in *She* magazine in March 2000.

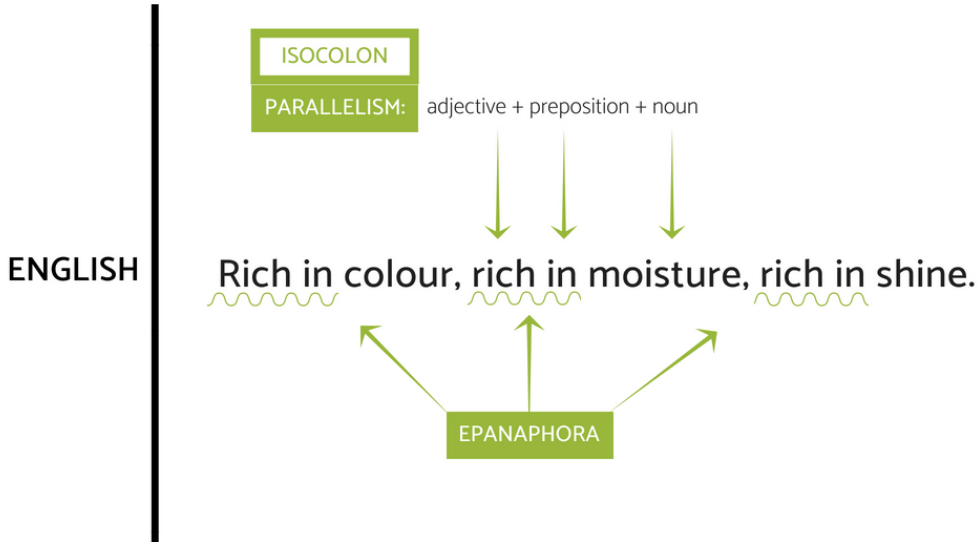

FIGURE 4.1: L'Oréal Paris Color Riche lipstick advertisement. *She* magazine, March 2000

In figure 4.1, we see two rhetorical figures in use. There is epanaphora in the repetition of rich at the beginning of each clause. The second device present in this example is the **adjective+preposition+noun** repetition known as isocolon[15].

Compare this with the same product's advertisement in the Russian edition of *Elle* magazine in July 2000 (figure 4.2). In this example, the epanaphora from the English version is not maintained in the three clauses. The final clause instead shows an example of polyptoton through the repetition of the root *roskoš* in both the noun and the adjective. The isocolon is not maintained exactly, but parallelism is marked in the three clauses.

[15]Isocolon is a series of similarly structured elements having the same length (*Silva Rhetoricae: The Forest of Rhetoric*)

FIGURE 4.2: L'Oréal Paris Color Riche lipstick advertisement. *Elle* magazine, July 2000

This demonstrates that often rhetorical figures need specific treatment when being translated - be it in a HT or MT process. In the absence of abundant studies on the the translation of rhetorical figures in MT, this may indicate that it is necessary to take each translation case on its own merit. One solution is to require a human in the loop approach when rhetorical figures are of interest in an MT process.

4.9 Interdisciplinarity and MT

By its nature, translation studies, including MT, is interdisciplinary and attracts researchers from a wide range of backgrounds (O'Brien and Saldanha, 2014; Odacıoğlu and Köktürk, 2015). Torres-Simón and Pym (2016)'s survey of 305 translation scholars found that people from a wide range of disciplinary backgrounds find their way into "working in Translation Studies or with translation or interpreting". These disciplinary backgrounds include language and linguistics (29%); teaching (20%) and business and finance (8%). Torres-Simón and Pym do not include MT in their study, but were it included under the translation studies umbrella, the disciplinary backgrounds would expand to include computer scientists, systems engineers and developers.

4.10 Summary

In this chapter, pertinent literature relating to rhetorical figures and MT was reviewed. The points put forward can be summarised as follows:

Rhetorical figures are inherent in arguably all forms of communication and they merit further exploration particularly in terms of their form and function and how they may be integrated into more technological scenarios. One of the key works which proposes an annotation scheme for rhetorical figures notes that in order to be able to utilise machine learning resources in a way that yields positive results, we need annotated texts which highlight rhetorical figures (Harris et al., 2018).

Machine translation is one of the most challenging natural language research areas, given that human translation itself is a difficult task. The nuances implicit in natural language present obstacles for machine translation systems, despite advances in the areas of SMT and NMT. There is potential to introduce linguistic domain expertise to MT in a grounded way, using exemplars and annotation.

Chapter 5

relo-KT Process Application

5.1 Introduction

Natural human speech is an intricate mélange of words, sounds, rhythms, schemes, tropes and many other combinations of linguistic tricks that we employ to fulfil a certain purpose when we communicate. In the case of political speech, rhetorical figures are often used to indirectly influence an audience to accept a politician's argument David (2014). It is precisely the intricacy and nuance of such devices that can make them challenging to detect using computational methods.

This chapter presents an in-depth case study, in which the relo-KT process outlined in Chapter 3 is applied to examine methods where linguistic understanding of rhetorical figures in political speech can be transferred to the MT domain.

"Case studies are a design of inquiry found in many fields, especially evaluation, in which the researcher develops an in-depth analysis of a case, often a program, event, activity, process, or one or more individuals" (Creswell, 2014). A case study "provides an opportunity for one aspect of a problem to be studied in some depth" (Bell, 2014). In order to determine whether the relo-KT process can be enacted, it was applied to a use case.

5.2 Applying the relo-KT process

The relo-KT process as depicted in Figure 5.1 and previously outlined in Chapter 3 is a process which was developed to formalise the knowledge transfer process between domains. This chapter explores whether the process works when it is applied to a use case concerning the domains which were discussed in Chapter 4, namely linguistics and MT.

For ease of reference, this will henceforth be referred to as the 'RF-MT' use case. RF is an abbreviation of rhetorical figures and MT of Machine Translation.

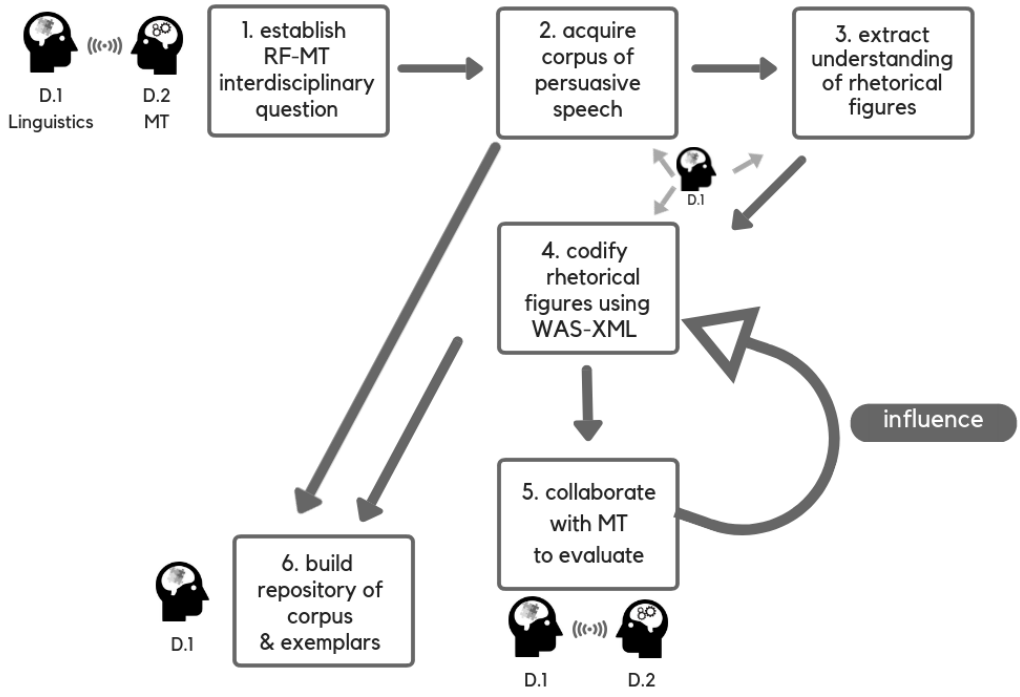

FIGURE 5.1: The relo-KT process for cross-disciplinary KT as applied in the RF-MT use case

Note about the domains

One of the key parts of the relo-KT process (Figure 5.1) is its iterative collaborative component which allows the insights derived through collaboration to feed back into the process to further fine tune it.

This research investigation commenced due to seeing an opportunity to influence machine processes with linguistic insight that was grounded in examples. Initially, the receiving domain (D.2) was Natural Language Generation (NLG). A first round of interviews was conducted to present a sample mark up schema for rhetorical figures and to gauge its practicality from NLG researchers. Influence drawn from the first round of interviews shaped the relo-KT process diagram in Figure 5.1.

5.2.1 Step 1: Question

For this use case, an interdisciplinary research question was identified in order to determine whether the relo-KT process could be enacted. As outlined in the previous section, the question initially focussed on the area of natural language generation (NLG). However, when Step 5 (collaborate to evaluate) of the process was enacted, it became clear through discussion and interaction that the NLG field is still at an early stage of development and integrating linguistic nuance (such as rhetorical figures) is not something which is a priority.

From the first iteration of interviews, the following question was established:

> to what extent can the rhetorical figures epanaphora, epistrophe, polyptoton and polysyndeton be identified from a corpus of political speech; and can the linguistic understanding of these figures then be transferred to the MT domain in order to aid MT processes such as post-editing?

To break down this question, there are two domains. The first is linguistics (D.1) which in this context encompasses elements of corpus linguistics

(the data) and discourse analysis (the rhetorical figures). The second domain (D.2) is MT.

5.2.2 Step 2: Acquire data

As we saw in Chapter 3, a linguistic corpus is a collection of texts which has been selected and compiled in order to study the language within them computationally (Wynne, 2005). Sinclair (2005) defines a corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research".

One of the aims of the RF-MT use case is to explore persuasive language. Political speech is persuasive, therefore the source for the data used in this study is the Official Report of Dáil Éireann[1]. The official report is "substantially but not strictly verbatim" (Oireachtas, 2018). This means that some light editing is carried out and the published record is not the exact words delivered in Dáil Éireann. This may be problematic for some types of corpus analysis. However, for this study, the statements are as they were intended to be delivered whether the TD[2] went off script or not. This is important because as outlined in Chapter 4, rhetorical figures tend to be carefully crafted and inserted in speech for a specific purpose. It is these linguistic features which are of interest in this study, therefore it is less important if they were spoken aloud and more important was the intent behind using them.

Pilot Study

A pilot study was carried out to assess whether the Official Report of Dáil Éireann was a suitable data source to work with for the overall study. The pilot study involved a close reading of key speeches from the 31st Dáil. These included emotive speeches and statements delivered on a variety of topics. These included an apology to clerical abuse victims on behalf of the State by

[1]The Official Report is a complete, authoritative and impartial written record of the proceedings of the Dáil, Seanad and Oireachtas committees: http://bit.ly/DE_OfficialReport
[2]TD is an abbreviation for Teachta Dála, a member of Dáil Éireann

Taoiseach[3], Enda Kenny and a statement on an austerity budget by Tánaiste, Joan Burton. The pilot study demonstrated that this type of corpus generates evidence of the type of rhetorical figures the relo-KT process hopes to transfer in a cross-disciplinary way.

Corpus

The criteria for inclusion in the corpus was that the statement had to be delivered on the same topic. The contentious issue of the 8th Amendment was debated in Dáil Éireann[4] in January 2018. It was a very emotive and polarised debate in which persuasive language was used. Contextual detail on the 8th Amendment and the debate is in Appendix A, and is summarised in the next paragraph and in Figure 5.2.

The text of these 63 statements, extracted from the Kildare Street website[5] makes up what I have termed the '8th Debates' corpus. The corpus consists of 99,630 words and can be accessed in XML format in this data repository: `https://dataverse.harvard.edu/dataverse/eclarkephd`. Figure 5.2 provides background information about the corpus.

[3]The Taoiseach is the head of government of Ireland and the Tánaiste is the deputy head of the government of Ireland
[4]Houses of the Oireachtas website: `https://www.oireachtas.ie/en/debates/find/`
[5]KildareStreet.com website: `https://www.kildarestreet.com/`

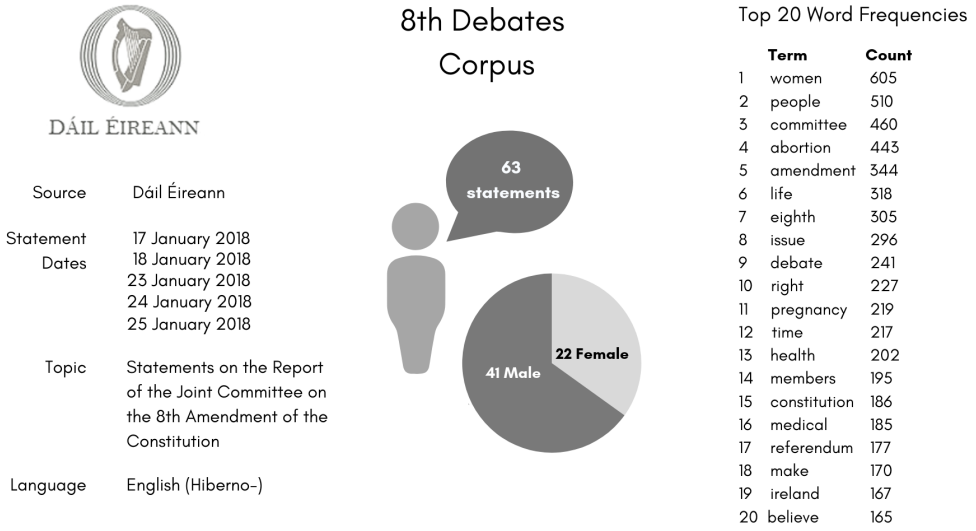

FIGURE 5.2: The 8th Debates corpus at a Glance

5.2.3 Step 3: Extract domain understanding

There are many rhetorical figures. Gideon Burton includes over four hundred of them on the aforementioned Silva Rhetoricae resource (*Silva Rhetoricae: The Forest of Rhetoric*). Their function tends to be subtle and they are often obscure, obsolete or esoteric. However, there are a number of common devices which form part of the persuasive speaker's repertoire. These devices often use repetition as the modus operandi and as a result, tend to follow certain patterns.

Pilot Study

In the pilot study stage of data extraction, two natural language processing methods were used to attempt to extract rhetorical figures from the corpus. Regular expressions are encoded text strings which can be used to match patterns in text. Take as an example the regular expression /b[aeiou]t/[6] which will match with "bat", "bet", "bit", "bot" and "but". This expression describes

[6]Example adapted from this website:
`https://lornajane.net/posts/2011/simple-regular-expressions-by-example`

anything which contains the letter "b" followed by a vowel and followed by "t". However, not only will it detect those three letter words, it will also detect the 'but' of '**but**ter' or the 'bet' of 'aided and a**bet**ted'. While regular expressions can be very useful for pattern-matching, they can become very complex very quickly.

This was the issue encountered when trying to use a regular expression to detect the tricolon[7] rhetorical figure. The most basic pattern of a tricolon is word, word, word as in the example **Veni, Vidi, Vici** which could be detected using this regular expression pattern: "\w+, \w+ and \w+". But what about the English translation of **Veni, Vidi, Vici**?

I came, I saw, I conquered would require a different pattern to detect it. Similar to the rule-based approach to MT (RBMT), this approach would require a list of patterns to potentially detect every type of tricolon.

Automated rhetorical figure detection

The key work relevant to this section is Java (2015)'s Rhetorica software. Rhetorica was developed to extract rhetorical devices from text. Rhetorica attempts to find and summarise a number of persuasive devices using the formalism for representing rhetorical devices laid out by Harris and DiMarco (2009) where possible.

Rhetorica

One of the goals which motivated Java's development of the Rhetorica software was: "to adopt and extend the automatic discovery of classical rhetorical figures described by Gawryjolek (2009)". Java then went on to test whether the figures discovered by Rhetorica could be used as "a discriminant in authorship attribution tasks". Stylometry is a technique which is

[7] A tricolon consists of three parallel clauses, phrases, or words, which happen to come in quick succession without any interruption: https://literarydevices.net/tricolon/

used to analyse such things as a text's authenticity or to identify the author of a particular text. Advances in technology and the use of computers for statistical analysis, coupled with a large corpora of digital text, mean that contemporary stylometry is a precise method with which text analysis can be used to identify patterns which can attribute genre, authorship or even gender of an author. Stylometric analysis techniques were famously used to establish the identity of the Unabomber terrorist (Koppel and Schler, 2003). Forensic analysis of the "Industrial Society and Its Future" manifesto demonstrated idiosyncratic writing characteristics which ultimately led to the Unabomber's capture and conviction (Goodman et al., 2007). Traditionally, stylometrics required close study of a small number of texts in order to find features in word usage or syntax which could attribute texts to a certain genre or author. However, with tools like Rhetorica, tasks which were previously painstakingly tedious, can now be carried out in a short space of time.

Rhetorica is a Windows command-line application for finding rhetorical figures in text[8]. The software was developed for English language text and can be used to identify up to 14 classical rhetorical figures, including the four that this case study is concerned with: epanaphora, epistrophe, polyptoton and polysyndeton.

Algorithm 1 detects figures of repetition in which word, phrase, clause, or grammatical structure patterns are repeated. Out of the four figures mentioned above, three can be classified as figures of repetition according to this algorithm: epanaphora, epistrophe and polysyndeton.

The fourth figure polyptoton is also repetition, but "of derivationally related forms of words" which "also allows for morphological similarity between words rather than just strict equality" as seen in the other three (Java, 2015). Rhetorica's Algorithm 2 "combines WordNet's synsets and lexical relations, and affix stemming, to find a word's derivational forms". Wordnet is "a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations." (*About WordNet* 2010). When searching a text for polyptoton,

[8]Rhetorica: `https://github.com/priscian/rhetorica`

Rhetorica ignores stop words[9].

In addition to WordNet, the Porter Stemming Algorithm[10] was used to remove common prefixes and suffixes from the words. Java notes that "The Porter stemmer does well with removing suffixes, less so with prefixes". Java attempted to improve the stemmer's performance "by checking the stem for common prefixes that might have escaped stemming" and concluded that "though the addition and removal of affixes to words— as well as the the Porter stemming itself—can create some false-positive related words, these do not adversely affect the algorithm results in any serious way".

Where possible, Rhetorica uses the formalism for representing rhetorical figures as laid out by Harris and DiMarco (2009). The context Java uses for rhetorical figure detection is phrases, clauses and sentences. The first step of the process is to find sentence boundaries within a text and Rhetorica uses the Stanford PCFG (probabilistic context-free grammar) Parser[11]. Once the sentences are detected, each one is "parsed, tokenized, and then broken into phrases and clauses derived from the parse tree" (Java, 2015). Java details how the parser works and outlines an example parse tree to demonstrate how the clauses and phrases are derived from sentences by the parser.

Java also notes that as "punctuation does not otherwise influence the discovery of rhetorical figures by our Rhetorica software", phrases and clauses derived from the parser are stored without any punctuation tokens.

[9]Stop words are the most common words in a language
[10]The Porter Stemming Algorithm: https://tartarus.org/martin/PorterStemmer/
[11]The Stanford Parser: A statistical parser:
https://nlp.stanford.edu/software/lex-parser.shtml

Element	Meaning
P	phrase
W	word
S	stem
M	morpheme
...	arbitrary intervening material*
...	morpheme boundaries
[...]	word boundaries
⟨...⟩	phrase or clause boundaries
$_{a, b, \ldots}$	identity $_{a = a}$, nonidentity $_{a \neq b}$

*Possibly null, with some upper limit; the shorthand is *proximal*

TABLE 5.1: Formalism for representing rhetorical figures (Adapted from Harris and DiMarco (2009) by Java (2015))

Java (2015) states that it performs the identification "with generally good precision and recall". Java defines precision and recall as:

Precision. The total number of examples of rhetorical figures correctly identified, divided by the total number of figures tested. Mathematically, the estimated precision (*prec*) is

$$prec = \frac{f_{++}}{f_{++} + f_{+-}} \tag{5.1}$$

where f_{++} is the total number of correctly identified figures (true positives) and f_{+-} is the total number of figures misidentified as the same figure (false positives). *High* precision, a measure of exactness, means that many more figures were correctly identified than misidentified.

Recall. The total number of examples of rhetorical figures correctly identified, divided by the total number of figures that should have been identified. Mathematically, the estimated recall (*rec*) is

$$rec = \frac{f_{++}}{f_{++} + f_{-+}} \tag{5.2}$$

where f_{++} is the total number of correctly identified figures (true positives) and f_{-+} is the total number of figures not identified as the same figure (false negatives). *High* recall, a measure of completeness, means that most of the figures were correctly identified.

In Chapter 4, examples of four rhetorical figures were presented. As part of the development of Rhetorica, each figure had to be formalised in order to identify them from a corpus of text. This is how Java formalised epanaphora, epistrophe, polyptoton and polysyndeton:

As we saw in Chapter 4, **epanaphora** is defined as the "repetition of the same word or group of words at the beginning of successive clauses, sentences, or lines" (*Silva Rhetoricae: The Forest of Rhetoric*). Java then formalises this as:

$$\begin{aligned} &\langle [W]_a \ldots \rangle \langle [W]_a \ldots \rangle \\ &\langle \langle P \rangle_a \ldots \rangle \langle \langle P \rangle_a \ldots \rangle \end{aligned} \tag{5.3}$$

We saw that **epistrophe** is "ending a series of lines, phrases, clauses, or sentences with the same word or words" (*Silva Rhetoricae: The Forest of Rhetoric*) which Java formalises as:

$$\begin{aligned} &\langle \ldots [W]_a \rangle \langle \ldots [W]_a \rangle \\ &\langle \ldots \langle P \rangle_a \rangle \langle \ldots \langle P \rangle_a \rangle \end{aligned} \tag{5.4}$$

Polyptoton is defined as the repetiton of "a word, but in a different form" This involves using a cognate[12] of a given word in close proximity:

[12] A cognate is a word which has the same linguistic derivation as another: https://en.oxforddictionaries.com/definition/cognate

$$\begin{aligned}
[S_a\,\{M_a\}]\,...\,[S_b\,\{M_b\}]\\
[\{M_a\}\,S_a]\,...\,[S_a\,\{M_b\}]\\
[S_a\,\{M_a\}]\,...\,[\{M_b\}\,S_a]\\
[\{M\}\,S_a\,\{M\}]\,...\,[S_a]\\
[S_a\,\{M\}]\,...\,[S_a]\\
etc.
\end{aligned} \tag{5.5}$$

For **Polysyndeton**, which is defined as "employing many conjunctions between clauses", Java formalises it as follows[13]:

$$and...and...and... \tag{5.6}$$

Java prepared a test file for each rhetorical figure. Each file contained at least 25 examples of that figure as culled from the Bible, literature, political speeches, popular culture, and common sayings and clichés. When possible, "the examples were left in the context of full sentences to more accurately simulate finding them *in situ*". Rhetorica yielded the results presented in Table 5.2.

According to Java (2015), the epanaphora test file had 28 examples of true epanaphora, and Rhetorica correctly identified all of them. Java notes that "to decrease the number of false-negative and incomplete anaphoras, Rhetorica ignores leading determiners, conjunctions, and prepositions in the comparison subsequences". The epistrophe test file had 42 examples of true epistrophe, and Rhetorica correctly identified all of them, but also identified two false positives. The polyptoton test file had 50 examples of true polyptoton, and Rhetorica correctly identified 45 of them, with 5 false negatives, and 2 false positives. Finally, Java's polysyndeton test file had 28 examples of true polysyndeton, and Rhetorica correctly identified all of them.

Ultimately, Java found that "classification models trained on Rhetorica's

[13] More than two repetitions are possible

Figure	Total No.	f^*_{++}	f^*_{+-}	f^*_{-+}	Misparse †	Prec. (%)	Recall (%)
Epanaphora	29	29	0	0	0(0)	100.0	100.0
Epistrophe	42	42	2	0	0(0)	95.0	100.0
Polyptoton	50	45	2	5	0(0)	96.0	90.0
Polysyndeton	28	28	0	0	0(0)	100.0	100.0

* f^*_{++}; true positive; f^*_{+-}: false positives; f^*_{-+}: false negatives.

† Total parser errors leading to false positives and negatives, with false negatives in parentheses.

TABLE 5.2: Precision and Recall Tests of the Rhetorica Software (Adapted from Java (2015))

rhetorical measures paired with lexical features typically performed better at authorship attribution than either set of features used individually". Given Java's success in identifying rhetorical figures with Rhetorica (as outlined above), it was decided that it was appropriate to use Rhetorica as a tool to detect rhetorical figures in the 8th Debates corpus.

The information outlined above about Rhetorica demonstrates that it is a tool which can extract information from a corpus which is linguistically meaningful. As previously mentioned, this study looks at four rhetorical figures from a potential fourteen that Rhetorica can identify. As outlined above, the four in question had good precision and recall. Not all of the fourteen performed as well in precision and recall tests which should be borne in mind if Rhetorica is the tool chosen to identify other devices, in particular tropes such as oxymoron.

Extracting domain understanding from the 8th Debates corpus

To extract domain understanding from the corpus, I adopted an approach which uses the Rhetorica software to identify four rhetorical figures in selected statements from the 8th Debates corpus. From the corpus, four statements were randomly selected for close reading. This involved a comparison of rhetorical figures detected by Rhetorica ("distant reading") with figures detected during the human reading ("close reading") of the speeches. The transcripts of the four statements are available in Appendix B and the complete

corpus is accessible in a Harvard Dataverse repository[14].

Statement	Word count	Rhetorical Figures
id01	2810	120
id02	2502	115
id04	1510	69
id41	1104	43
Totals	7926	347

TABLE 5.3: Word counts and rhetorical figure totals

A close reading and manual mark up was required for each statement in the sample in order to determine whether Rhetorica's figure detection was appropriate. The four statements combined amount to 7926 words. A total of 347 figures were identified by Rhetorica (Table 5.3). Close reading is a laborious and time-consuming task. However, it was deemed that a close reading of four statements could be completed in a manageable amount of time. If more statements were needed to get more illustrative examples, they could be added, but for each added statement, there is an additional time cost in terms of the human validation of the Rhetorica software's output.

Firstly, the plain text files were run through the Rhetorica software which produced two output files per statement (see Appendix C for complete output files) which I renamed as 'id_POS' (Part of Speech) and 'id_figures'. The output files contain the data in Tables 5.4 (id_POS) and 5.5 (id_figures).

The files required some light post-processing which involved removing data relating to rhetorical figures which are not relevant to this study.

[14]Harvard Dataverse respository:
https://dataverse.harvard.edu/dataverse/eclarkephd

id_POS	
sentence_id	assigns a unique id to each sentence
token_id	assigns an id to each individual token in the sentence
word	the word used in the sentence
left_edge	the character at which the word begins
right_edge	the character at which the word ends
tag	assigns a tag from the Penn Treebank Part of Speech (POS) Tag Set (Appendix D)
tag_equiv	Gawryjolek (2009) (upon whose work Java (2015)'s expands) used broader equivalence classes to represent major parts of speech than the Penn Treebank POS (this is of importance for some of the figures of speech which Rhetorica detects (e.g. isocolon), but it is not relevant to this work)
depth	refers to the depth of the parse tree (this is of importance for some of the figures of speech which Rhetorica detects (e.g. isocolon), but it is not relevant to this work)
stem	the stem of the word with the more common morphological and inflectional endings removed

TABLE 5.4: Natural Language Processing (NLP) data contained in the id_POS output file

To demonstrate how the two Rhetorica output files correspond to each other, take as an example sentence 25 from the 8th Debates statement id04 (represented as Rhetorica output in figure 5.3).

"This obsessive control of women did not happen by accident."

This is the twenty-sixth sentence in statement id04, so it has been assigned sentence id number 25 (the first sentence is 0). 'This' is the first token, 'obsessive' is the second, 'control' is the third and so on. The first token 'This' begins at character 0 (left edge) and ends at character 4 (right edge). The

sentence_id	token_id	word	left_edge	right_edge	tag	tag_equiv	depth	stem
25	0	This	0	4	DT	DT	3	Thi
25	1	obsessive	4	13	JJ	JJ	3	obsess
25	2	control	13	20	NN	NN	3	control
25	3	of	20	22	IN	IN	3	of
25	4	women	22	27	NNS	NN	2	women
25	5	did	27	30	VBD	VB	4	did
25	6	not	30	33	RB	RB	4	not
25	7	happen	33	39	VB	VB	3	happen
25	8	by	39	41	IN	IN	2	by
25	9	accident	41	49	NN	NN	1	accid
25	10	.	49	50	.	.	5	.

FIGURE 5.3: Rhetorica 'id_POS' output for sentence 25 from 8th Debates statement id04

second token 'obsessive' begins at 4 and ends at 13 and so on.

The id_figures table (5.5), on the other hand, contains data relating directly to the rhetorical figures detected by Rhetorica.

Figure 5.4 shows the total output for the rhetorical figure epistrophe in statement id04. Rhetorica detected nine occurrences of epistrophe in total (figures 111-119). To explore an example from statement id04 further, Epistrophe 113 corresponds to the output seen in figure 5.3 previously.

> This obsessive control of women **did not happen by accident**.
> It was very much intended by a powerful conservative cohort across society - in government and the churches, across the highest ranks of the public and Civil Service and among the professional elites.
> Women's subjugation was part of a carving up of power and influence in the public and private spheres, which **did not happen by accident**.

(Statement id04, 17 January 2018)

In-depth findings from this stage of the process will be presented in Chapter 6 as part of the evaluation of the relo-KT process. However, a number of examples from the statements delivered during the 8th Debates are outlined below as they are important to demonstrate the knowledge codification stage of the relo-KT process (step 4, section 5.2.4).

id_figures	
figure_id	assigns a unique id to each rhetorical figure detected
token_id	assigns a sequential number to each individual token in the sentence
type	refers to the figure of speech
word	the words which make up the figure of speech
sentence_id	refers to the unique id assigned to each sentence in the POS output
left_edge	the character at which the word begins
right_edge	the character at which the word ends
tag	assigns a tag from the Penn Treebank Part of Speech (POS) Tag Set (Appendix D)
tag_equiv	Gawryjolek (2009) (upon whose work Java (2015)'s expands) used broader equivalence classes to represent major parts of speech than the Penn Treebank POS (this is of importance for some of the figures of speech which Rhetorica detects (e.g. isocolon), but it is not relevant to this work)
depth	refers to the depth of the parse tree (this is of importance for some of the figures of speech which Rhetorica detects (e.g. isocolon), but it is not relevant to this work)
stem	the stem of the word with the more common morphological and inflectional endings removed

TABLE 5.5: Natural Language Processing (NLP) data contained in the id_figures output file

figure_id	token_id	type	word	sentence_id	left_edge	right_edge	tag	tag_equiv	depth	stem
111	0	Epistrophe	the	2	25	28	DT	DT	5	the
111	1	Epistrophe	committee	2	28	37	NN	NN	5	committe
111	2	Epistrophe	the	3	53	56	DT	DT	2	the
111	3	Epistrophe	committee	3	56	65	NN	NN	2	committe
112	0	Epistrophe	work	6	75	79	NN	NN	1	work
112	1	Epistrophe	work	7	128	132	NN	NN	2	work
113	0	Epistrophe	did	25	27	30	VBD	VB	4	did
113	1	Epistrophe	not	25	30	33	RB	RB	4	not
113	2	Epistrophe	happen	25	33	39	VB	VB	3	happen
113	3	Epistrophe	by	25	39	41	IN	IN	2	by
113	4	Epistrophe	accident	25	41	49	NN	NN	1	accid
113	5	Epistrophe	did	27	90	93	VBD	VB	4	did
113	6	Epistrophe	not	27	93	96	RB	RB	4	not
113	7	Epistrophe	happen	27	96	102	VB	VB	3	happen
113	8	Epistrophe	by	27	102	104	IN	IN	2	by
113	9	Epistrophe	accident	27	104	112	NN	NN	1	accid
114	0	Epistrophe	of	35	29	31	IN	IN	10	of
114	1	Epistrophe	rape	35	31	35	NN	NN	9	rape
114	2	Epistrophe	of	36	100	102	IN	IN	2	of
114	3	Epistrophe	rape	36	102	106	NN	NN	1	rape
115	0	Epistrophe	the	38	151	154	DT	DT	1	the
115	1	Epistrophe	X	38	154	155	NNP	NN	1	X
115	2	Epistrophe	case	38	155	159	NN	NN	1	case
115	3	Epistrophe	the	40	164	167	DT	DT	1	the
115	4	Epistrophe	X	40	167	168	NNP	NN	1	X
115	5	Epistrophe	case	40	168	172	NN	NN	1	case
116	0	Epistrophe	issue	47	45	50	NN	NN	1	issu
116	1	Epistrophe	issue	48	127	132	NN	NN	3	issu
117	0	Epistrophe	health	49	54	60	NN	NN	8	health
117	1	Epistrophe	health	49	68	74	NN	NN	8	health
118	0	Epistrophe	the	50	47	50	DT	DT	3	the
118	1	Epistrophe	joint	50	50	55	JJ	JJ	3	joint
118	2	Epistrophe	committee	50	55	64	NN	NN	3	committe
118	3	Epistrophe	the	52	66	69	DT	DT	1	the
118	4	Epistrophe	joint	52	69	74	JJ	JJ	1	joint
118	5	Epistrophe	committee	52	74	83	NN	NN	1	committe
119	0	Epistrophe	the	52	35	38	DT	DT	2	the
119	1	Epistrophe	Constitution	52	38	50	NNP	NN	2	Constitut
119	2	Epistrophe	the	54	137	140	DT	DT	1	the
119	3	Epistrophe	Constitution	54	140	152	NNP	NN	1	Constitut

FIGURE 5.4: Rhetorica output for epistrophe detected in 8th Debates statement id04

1. **Epanaphora**
   repetition of the same word or group of words at the beginning of successive clauses, sentences, or lines

   > **It is important** we all recognise that fundamentally this is a deeply personal issue, and it is a matter for each individual citizen, as part of a referendum, to decide what he or she believes is right. **It is important** that all of us are non-judgmental in this regard and that we respect the views and opinions of others.

   (Statement id41, 23 January 2018)

2. **Epistrophe**
ending a series of lines, phrases, clauses, or sentences with the same word or words

> I cannot help but wonder what we would have done if we did not have a neighbouring island to help us turn **a blind eye**. Sometimes turning **a blind eye** is the same as turning your back.

(Statement id01, 17 January 2018)

3. **Polyptoton**
repeating a word, but in a different form. Using a cognate of a given word in close proximity

> If we are to allow the people the space to **grapple** with this issue, we in this Chamber must equally acknowledge that people have deeply held views - morally, ethically and even religiously.
> That aside, my personal or political discomfort is nothing compared with the discomfort caused to women every day of the week who are **grappling** with crisis or unwanted pregnancies...

(Statement id02, 17 January 2018)

4. **Polysyndeton**
employing many conjunctions between clauses

> It is now time for the law, politics and every Member of the Oireachtas to catch up with public opinion **and** the new Ireland, the country in which my 14 year old daughter **and** all our daughters **and** granddaughters - **and** our boys **and** men - live **and** give us and them a decent, human rights-based **and** respectful Constitution that acknowledges women as full **and** equal people.

(Statement id04, 17 January 2018)

5.2.4 Step 4: Knowledge codification

This section will show how specific understanding from one domain (D.1) (e.g. linguistic understanding of Rhetorica's output) will be codified using a knowledge lens to produce information that can be consumed by a practitioner from the receiving domain (D.2) (MT).

As presented in Chapter 4, one of the main challenges of conveying an understanding of rhetorical figures in machine learning tasks such as MT is addressing the "bottleneck" created by "the lack of annotated data" containing rhetorical figures (Dubremetz, 2017). Harris et al. (2018)'s attempt to meet this challenge is to offer an XML annotation scheme known as the Waterloo Annotation Schema for Rhetorical Figures (WAS) which they believe "holds considerable promise" when it comes to "computational understanding of natural language".

The WAS annotation scheme was applied to Rhetorica's output to annotate the rhetorical figures using standoff mark up. Standoff markup is placed outside of the text it is meant to tag. In the RF-MT use case, the mark up is in a separate XML file. This is standard procedure when overlap occurs in a marked up text.

This is how the WAS annotation schema was applied to the rhetorical figures presented in section 5.2.3.

1. **Epanaphora**

> **It is important** we all recognise that fundamentally this is a deeply personal issue, and it is a matter for each individual citizen, as part of a referendum, to decide what he or she believes is right.
> **It is important** that all of us are non-judgmental in this regard and that we respect the views and opinions of others.

(Statement id41, 23 January 2018)

```
<rhetorical_figure name="epanaphora" figure_id="65">

  <occurrences>

    <occurrence id="1" sentence_id="7" element="it is important">
      <left_edge>0</left_edge>
      <right_edge>13</right_edge>
    </occurrence>

    <occurrence id="2" sentence_id="8" element="It is important">
      <left_edge>0</left_edge>
      <right_edge>13</right_edge>
    </occurrence>

  </occurrences>

</rhetorical_figure>
```

2. **Epistrophe**

> I cannot help but wonder what we would have done if we did not have a neighbouring island to help us turn **a blind eye**.
> Sometimes turning **a blind eye** is the same as turning your back.

(Statement id01, 17 January 2018)

```
<rhetorical_figure name="epistrophe" figure_id="222">

  <occurrences>

    <occurrence id="1" sentence_id="20" element="a blind eye">
      <left_edge>84</left_edge>
      <right_edge>93</right_edge>
    </occurrence>

    <occurrence id="2" sentence_id="21" element="a blind eye">
      <left_edge>16</left_edge>
      <right_edge>25</right_edge>
    </occurrence>

```

```
  </occurrences>

</rhetorical_figure>
```

3. **Polyptoton**

> If we are to allow the people the space to **grapple** with this issue, we in this Chamber must equally acknowledge that people have deeply held views - morally, ethically and even religiously.
> That aside, my personal or political discomfort is nothing compared with the discomfort caused to women every day of the week who are **grappling** with crisis or unwanted pregnancies...

(Statement id02, 17 January 2018)

```
<rhetorical_figure name="polyptoton" figure_id="209">

  <occurrences>

    <occurrence id="1" sentence_id="23" element="grapple">
      <left_edge>33</left_edge>
      <right_edge>40</right_edge>
    </occurrence>

    <occurrence id="2" sentence_id="24" element="grappling">
      <left_edge>111</left_edge>
      <right_edge>120</right_edge>
    </occurrence>

  </occurrences>

</rhetorical_figure>
```

4. **Polysyndeton**

> It is now time for the law, politics and every Member of the Oireachtas to catch up with public opinion **and** the new Ireland, the country in which my 14 year old daughter **and** all

> our daughters **and** granddaughters - **and** our boys **and** men - live **and** give us and them a decent, human rights-based **and** respectful Constitution that acknowledges women as full **and** equal people.

(Statement id04, 17 January 2018)

```
<rhetorical_figure name="polysyndeton" figure_id="84">

  <occurrences>

    <occurrence id="1" sentence_id="68" element="and">
      <left_edge>29</left_edge>
      <right_edge>32</right_edge>
    </occurrence>

    <occurrence id="2" sentence_id="68" element="and">
      <left_edge>84</left_edge>
      <right_edge>87</right_edge>
    </occurrence>

    <occurrence id="3" sentence_id="68" element="and">
      <left_edge>137</left_edge>
      <right_edge>140</right_edge>
    </occurrence>

    <occurrence id="4" sentence_id="68" element="and">
      <left_edge>155</left_edge>
      <right_edge>158</right_edge>
    </occurrence>

    <occurrence id="5" sentence_id="68" element="and">
      <left_edge>173</left_edge>
      <right_edge>176</right_edge>
    </occurrence>

    <occurrence id="6" sentence_id="68" element="and">
      <left_edge>183</left_edge>
      <right_edge>186</right_edge>
    </occurrence>

    <occurrence id="7" sentence_id="68" element="and">
      <left_edge>194</left_edge>
      <right_edge>197</right_edge>
    </occurrence>
```

```
39
40    <occurrence id="8" sentence_id="68" element="and">
41      <left_edge>203</left_edge>
42      <right_edge>206</right_edge>
43    </occurrence>
44
45    <occurrence id="9" sentence_id="68" element="and">
46      <left_edge>235</left_edge>
47      <right_edge>238</right_edge>
48    </occurrence>
49
50    <occurrence id="10" sentence_id="68" element="and">
51      <left_edge>287</left_edge>
52      <right_edge>290</right_edge>
53    </occurrence>
54
55  </occurrences>
56
57 </rhetorical_figure>
```

The codification is in XML format which means that can be accepted by MT systems. This codification was developed collaboratively between D.1 and D.2 practitioners. The collaboration required to reach this codification is discussed in more detail in the following section.

5.2.5 Step 5: Collaborate to evaluate

The previous four steps of this process have led to the creation of a corpus of political speech from which a bank of annotated rhetorical figures has been assembled. The purpose of the previous steps has been to go towards developing a process for the transfer of linguistic understanding of rhetorical figures to ML/MT domains.

Interviews

Semi-structured interviews were chosen as a mode to determine whether the mark up of rhetorical figures in parliamentary speech is of value to practitioners in the MT domain. The interviews carried out in the course of this study were semi-structured and were used to gain detailed insight into interviewees' understanding of how rhetorical figures and persuasive language are currently handled by MT systems and where they see future developments of the field. The interviews were conducted along a set of guiding questions, and took the form of a conversation rather than a rigid question and answer format. Each interview was recorded using audio recording software on a Macbook Pro (namely Quicktime[15]).

Following each interview, the recording was transcribed into written format, codified and analysed using the ATLAS.ti software[16]. ATLAS.ti is a "workbench for the qualitative analysis of large bodies of textual" data. Creswell (2014) outlines the benefits of using a qualitative software program like ATLAS.ti stating that software programs are a "logical choice for qualitative data analysis over hand coding". Among the benefits, Creswell notes the efficiency of searching qualitative data using a computer program meaning that a "researcher can quickly locate all passages coded the same". Additionally, Creswell highlights that qualitative software can facilitate the comparison and visualisation of differently coded data.

The full transcripts of the interviews are in Appendices G and H. The recordings and their transcriptions have been fully anonymised and are stored digitally in accordance with the Data Protection Act at Trinity College, Dublin. Each interview lasted between a half an hour to 1 hour depending on the participant's availability and the development of the interview. Prior to the interview, each participant was provided with a copy of the informed consent form and participant information which they could consult to familiarise themselves with the interview process. These are available in Appendices E and F.

[15]Quicktime: https://support.apple.com/quicktime
[16]ATLAS.ti: https://atlasti.com/

The interview cycle was designed to be an iterative process. Each cycle of interviews, and indeed responses from each interview, was fed back into the process itself.

First round of interviews

The first iteration aimed to explore how the understanding of human linguistic nuance in the form of rhetorical figures used to persuade can be 'translated' or understood from the linguistics domain to the domain of Natural Language Generation (NLG). In brief, how can the human linguistic understanding of how these patterns work be presented to practitioners within the NLG community in order that it could be leveraged and implemented in their systems.

From online biographies of ADAPT Centre researchers, potential interviewees were identified and approached over a time period from May 2017-August 2017. In total, four interviews were carried out with researchers who work in the NLG field. Their profiles at the time of interview were:

1. Research fellow in Speech Processing
2. PhD student in Speech Communications
3. Research fellow in Computer Science
4. Research fellow in Machine Translation, Information Retrieval and Machine Learning

The outcome of the first round of interviews can be summarised as follows:

- Trends in NLG are moving from a rule-based approach towards methods which use statistical deep learning algorithms
- There is value in capturing the nuance of rhetorical figures
- Disambiguation will present a challenge
- Some form of XML mark up would be useful for incorporating rhetorical figures into machine based systems

From the interview process, it emerged that NLG is a very broad and diverse area to impact on the whole. Collectively, the interviewees identified as NLG researchers, but on an individual basis, their research sits in areas such as speech synthesis, computational linguistics, dialogue systems and MT.

It also became clear that machine translation MT is an area in which the retention of linguistic features like rhetorical figures is useful. Some of the comments from the interview with the MT specialist include the significance of contextual information about a particular figure; and the importance of retaining its persuasive purpose in the source text when it is translated in the target text.

> It is very important for an MT system to keep this [epanaphora] in the target language. But, in the real MT systems we cannot guarantee that this is kept in the translation because of the word order
> *Round 1, Interviewee 4*

The MT interviewee used the following example of epanaphora from Charles Dickens' A Tale of Two Cities to demonstrate how rhetorical figures may be handled by an MT system:

> "**It was the** best of times, **it was the** worst of times, **it was the** age of wisdom, **it was the** age of foolishness, **it was the** epoch of belief, **it was the** epoch of incredulity, **it was the** season of light, **it was the** season of darkness, **it was the** spring of hope, **it was the** winter of despair."

In an MT system, there is no guarantee that the 'it was the' epanaphora will be translated consistently in the target language. It might be 'it was' in the first occurrence, but in a subsequent occurrence, it could be changed to 'it has been' for example. The mark up I am suggesting in this work, could then be used as a feature in the MT system, so that a pattern could be recognised as a particular figure, which could then be highlighted in the source sentence and thus retained in the target sentence.

> if we have this markup, then we can use this at least as a feature in the MT system. That the system can recognise, ok this is for example epanaphora, and then we can keep the format in the source sentence and then keep this in the target sentence.
> *Round 1, Interviewee 4*

Second round of interviews

The second round of interviews focused solely on MT researchers. Round one indicated that NLG researchers believe a deeper understanding of rhetorical figures is of value in their field, but for the most part, they were unable to provide a concrete example of how this research might be incorporated in their systems. The MT researcher on the other hand demonstrated clearly how the subtleties of these figures might literally be lost in translation.

The desired outcome from the second round of interviews was:

- to gain a deeper understanding of the MT process and the role of the human in the process – be it in pre- or post-editing
- to determine whether rhetorical figures (or similar linguistic devices) have a tendency to be lost in the translation process
- to understand how rhetorical figures could be represented in a markup schema which could be of use within an MT system

Four interviewees participated in round two. Their backgrounds at the time of interviewing were:

1. PhD student in MT
2. PhD student in MT, Languages and Linguistics
3. Research fellow in machine and human evaluation of MT
4. Research fellow in MT

The questions in table 5.6 were prepared, but as the interview was semi-structured, it was not restricted to these alone. Due to the organic nature of semi-structured interviewing, the exact line of questioning differed slightly from interview to interview depending on how the conversation branched.

Detailed findings from the second round of interviews are presented and discussed in Chapter 6.

5.2.6 Step 6: Build repository

A small dataset has been developed as part of this research. This has been shared on Harvard's Dataverse[17] and follows the Austin Principles of Data Citation in Linguistics[18] which recognise "the dual necessity of creating citation practices that are both human understandable and machine-actionable" (Berez-Kroeker et al., 2018).

The dataset materials related to this thesis can be located at this URL: `https://dataverse.harvard.edu/dataverse/eclarkephd`.

5.3 Summary

In Chapter 5, the relo-KT process was applied to a use case which takes linguistic understanding of rhetorical figures and transfers them to the MT domain. In order to do this, a corpus of political statements delivered in Dáil Éireann was created. Then, Java (2015)'s Rhetorica software was used to extract occurrences of four rhetorical figures from the corpus. Linguistic understanding of exemplar rhetorical figures for persuasion was codified using Rhetorica's output and the WAS-XML annotation schema. In an iterative cycle of semi-structured interviews, I collaborated with NLG and MT experts to define a codification schema which is compatible with both domains. A detailed discussion of the findings from this RF-MT use case follows in Chapter 6.

[17] Harvard Dataverse repository:
`https://dataverse.harvard.edu/dataverse/eclarkephd`
[18] Austin Principles of Data Citation in Linguistics:
`https://site.uit.no/linguisticsdatacitation/austinprinciples/`

General
Could you describe your own work? Does any of your work deal with persuasive speech? (or similar – e.g. sarcasm / humour / metaphor) Is pre- or post-editing a part of your process? What role does the human have in your system?
MT and human-editing
What are the qualities required of an MT pre- or post-editor?
MT and rhetoric
What is the current state of capturing nuance like rhetorical figures in MT? How is language like rhetorical figures retained through the translation process? Does this differ depending on whether it is a machine or a human translation?
Value of proposed approach
How should rhetorical figures be marked up in order for them to be useful? How wide does the markup need to be in order to be useful? How might this markup be incorporated in a system you work with? How might this approach be validated in the translation domain?

TABLE 5.6: Interview questions

Chapter 6

Analysis and discussion

6.1 Introduction

In the introduction to this thesis (Chapter 1), two research questions were posed. RQ1 relates to formalising a holistic method of cross-disciplinary knowledge transfer which until now had been missing. The relo-KT process has been developed to provide a structured, collaborative approach to cross-disicplinary knowledge transfer through knowledge codification and interdisciplinary interaction.

To answer RQ2, the relo-KT process was enacted in the RF-MT use case (rhetorical figure-machine translation) to demonstrate the value it can bring to an interdisciplinary research scenario. Grounding D.1 knowledge of rhetorical figures in marked up examples, and using the exemplars as artefacts in interviews with D.2 experts, provides a meaningful structure for transferring D.1 understanding to D.2. The results from the enactment of the process, through the specific RF-MT use case, demonstrate its effectiveness. These results will be presented and discussed in this chapter.

This chapter adopts a two-pronged approach to collate the various findings of the subject study. Firstly, each distinct step of the relo-KT process for cross-disciplinary knowledge transfer is analysed and discussed, to demonstrate the value it brings to the knowledge transfer (henceforth KT) process. These findings are based on the RF-MT use case which explored how the relo-KT process might be enacted to transfer knowledge from the linguistics domain to the MT domain. Secondly, a critical analysis of the enactment of the relo-KT process is presented.

6.2 Overview

The literature review in Chapter 2 revealed that while there are guides to doing interdisciplinary research, such as Szostak (2002)'s twelve step process for doing interdisciplinarity and the PMI (2015)'s knowledge transfer life cycle, there lacked a formalised method for transferring knowledge between disciplines.

An exploration into how cross-disciplinary KT is performed in five interdisciplinary research projects indicated that collaboration was clearly involved in creating the associated outputs and resources. However, none of the projects examined have explicitly documented the collaborative processes used, beyond referring to the disciplinary backgrounds of the project participants. This is indicative of collaborative elements in many research and work scenarios in which interactions and discussions between stakeholders tend to be documented in meeting minutes or internal documents. In general, these tend not to be publicly accessible.

From the literature presented in Chapter 4, it emerged that research on rhetorical figures and MT merits further exploration. Smith (2006)'s work on English-Russian translations was particularly informative and demonstrated the levels of nuance which must be considered when translating persuasive linguistic devices. Beyond Smith's work, there was little to indicate how MT systems might integrate such figures, ergo exposing a gap.

One of the potential reasons for this could be that MT is a rapidly evolving field. This is demonstrable by considering that when work began on this PhD in 2014, statistical MT was de rigueur. Within two years, the entire field had shifted towards neural MT which has become the current paradigm in MT systems (Castilho et al., 2017). One of the consequences of being in a field which is advancing quickly is that the focus is on developing and improving the systems. As a result, adapting and customising systems to integrate or recognise nuanced features such as rhetorical figures is seen as a future endeavour (*The State of Neural Machine Translation (NMT)*).

6.3 Evaluation of relo-KT

The relo-KT process (which is reproduced in Figure 6.1 to aid memory) brings together multiple steps which facilitate interdisciplinary collaboration. The process had a number of key advantages which will be explored in this section.

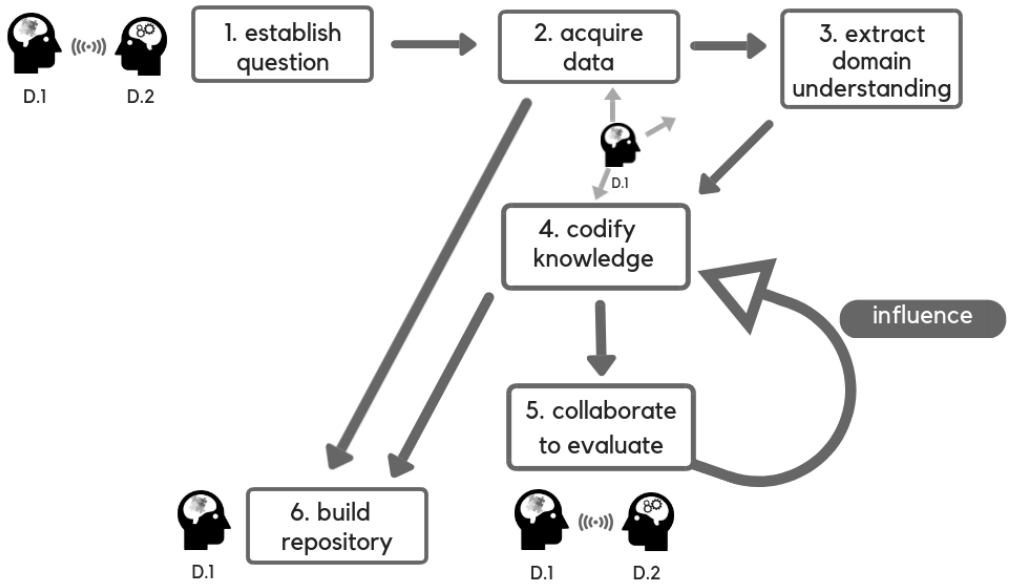

FIGURE 6.1: The relo-KT process for cross-disciplinary KT as applied in the RF-MT use case

In summary, the relo-KT process proved successful at cross-disciplinary KT because it enabled D.1 knowledge to be codified into examples which could be articulated and demonstrated to D.2 experts through semi-structured interviews. It also allowed the refinement of the codified data, in order for it to be readily interpreted by MT software and platforms. The D.1-D.2 collaboration, which took place in the RF-MT use case, demonstrated that integrating rhetorical figure understanding is a challenge for MT researchers. The collaboration further showed that the challenge is worth addressing and through

iterations of the process, an understanding of the impact this could have in a broader MT system was developed.

The subsequent sections analyse and evaluate each step of the relo-KT process as it was enacted in the RF-MT use case.

Step 1: Establish question

Establishing an interdisciplinary question involved interaction with practitioners in D.2, which as seen, did not start out as MT[1]. Through cross-disciplinary interaction and collaboration it became clear that the MT domain is primed to focus on integrating linguistic nuance. Thus, the interdisciplinary question emerged.

The question establishment step indicates the iterative nature of the relo-KT process. The process also reflects that interdisciplinary research is rarely a straightforward route from A to B. While certain steps of the relo-KT process depend on each other, I do not prescribe that the flow presented must be followed strictly and depending on the nature of the interdisciplinary enquiry certain steps may be visited and revisited in different sequences and iterations.

Step 2: Acquire data

The pilot study was a useful exercise to undertake at the beginning of the data acquisition stage, as it indicated that rhetorical figures could be identified within the intended corpus. Based on the findings of the pilot study, I gathered a corpus of Dáil Éireann statements called the '8th Debates' corpus for exploration in the RF-MT use case.

There are numerous benefits to using a corpus of Dáil statements for linguistic analysis. For example, the debates are available in electronic format under

[1]Over the course of the first round of interviews, RF-MT question evolved to: "To what extent can the rhetorical figures epanaphora, epistrophe, polyptoton and polysyndeton be identified from a corpus of political speech; and can the linguistic understanding of these figures then be transferred to the MT domain in order to aid MT processes such as post-editing?"

an Oireachtas (Open Data) PSI Licence[2] which waives any requirement under the Re-use of Public Sector Information Regulations to formally apply for permission to re-use information covered by it. As the data is in the public domain, permission was not required to use the statements for textual analysis. In the context of the RF-MT use case, which explores persuasive language, political statements are an obvious choice due to politicians' tendency to use language to influence.

As the subject matter of the 8th Debates was emotive and polarising, it can be assumed that any TD who chose to speak on the record about the topic would have crafted their statement carefully in order to avoid misinterpretation or obfuscation. Political speeches tend to be carefully crafted in order to fulfil a purpose, and speechwriters typically draw upon a range of techniques which include rhetorical figures to achieve persuasion (Atkinson, 2004). Thus, the use of the 8th Debates corpus for exploring rhetorical figure usage in political speech is justified.

Step 3: Extract domain understanding

The Rhetorica software was chosen as a tool for the task of extracting four rhetorical figures from the corpus based on Java (2015)'s successful detection of rhetorical figures with it, as outlined in Chapter 5.

On the whole, Rhetorica is a user friendly tool. Java has provided clear example-based documentation for users to follow. The input text does not require a lot of bespoke pre-processing to ensure it is Rhetorica-readable, provided it is plain text. The output files are easy for a machine to understand, and also easy for a human to decipher.

A drawback of Rhetorica however is that it is a Windows command-line application[3], so it is restrictive for users of other operating systems.

To fully understand what Rhetorica's output meant in terms of the 8th Debates corpus, I followed Baker (2006)'s suggestion that familiarisation with

[2]Oireachtas (Open Data) PSI Licence: https://www.oireachtas.ie/en/open-data/
[3]Rhetorica: https://github.com/priscian/rhetorica

the corpus is a practical solution to ensure that the "analyst does not commence from the position of *tabula rasa*". For this reason, I conducted a close reading of four statements randomly selected from the 8th Debates corpus and compared them with Rhetorica's raw output. The length (words) of the four statements is in table 6.1.

Statement	Length (words)
id01	2810
id02	2502
id04	1510
id41	1104

TABLE 6.1: Statement word counts

Rhetorica detected the rhetorical figures in table 6.2 from four statements.

Statement	Epanaphora	Epistrophe	Polyptoton	Polysyndeton
id01	60	10	42	8
id02	51	14	45	5
id04	25	9	26	9
id41	22	11	10	0

TABLE 6.2: Rhetorica output

Epanaphora and Epistrophe

Rhetorica can be overzealous in its epanaphora and epistrophe figure detection. An example from the 8th Debates is the following detection of

epanaphora from statement id02. As previously seen, epanaphora is defined as the "repetition of the same word or group of words at the beginning of successive clauses, sentences, or lines" (*Silva Rhetoricae: The Forest of Rhetoric*). However in the example below, Rhetorica detects a false positive epanaphora of 'to' across six sentences. From a human reading, it is clear that this is unlikely to be an intended epanaphora.

> It is our duty **to** address this issue. With the best will in the world, and although people have varying opinions on this, we cannot do anything other than what we have legislated for already unless we repeal, amend or replace Article 40.3.3°. We have **to** change what is in our Constitution. That will be the first step towards addressing this issue.
>
> It would be easy for me and others contributing in this debate **to** keep our heads down and hope that the issue goes away but, as a generation of politicians, we have **to** deal with it. We have **to** give the people an opportunity **to** express their opinion in light of the fact that the last time they had such an opportunity was in 1983.
>
> *Epanaphora 137, 8th Debates statement id02*

This demonstrates that human intervention is required to post-process the Rhetorica output to remove noise. In Natural Language Processing NLP, a token is "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing" (*Tokenization*). Tokens are often loosely referred to as words. The approach taken in this study was to remove all one-token occurrences of epanaphora and epistrophe (such as Epanaphora 137) when post-processing Rhetorica's output which resulted in tables 6.3 and 6.4.

Statement	Epanaphora	Epanaphora 2+ tokens
id01	60	28
id02	51	15
id04	25	7
id41	22	11

TABLE 6.3: Post-processed Epanaphora

Statement	Epistrophe	Epistrophe 2+ tokens
id01	10	4
id02	14	9
id04	9	6
id41	11	4

TABLE 6.4: Post-processed Epistrophe

Once the one-token epanaphora and epistrophe occurrences are removed, the numbers are significantly lower.

Polyptoton

Generally, Rhetorica handled polyptoton well and presented some examples which a human reader may not register when reading the statements. An example of this is the use of the cognates grapple/grappling in Statement id02 which spans two paragraphs.

> If we are to allow the people the space to **grapple** with this issue, we in this Chamber must equally acknowledge that people have deeply held views - morally, ethically and even religiously.
>
> That aside, my personal or political discomfort is nothing compared with the discomfort caused to women every day of the week who are **grappling** with crisis or unwanted pregnancies ...

Polyptoton 209, 8th Debates statement id02

However, there were occurrences when polyptoton presented more of a challenge for Rhetorica (Table 6.5) which I would like to explore in some detail.

Statement	Polyptoton	Polyptoton not cognates
id01	42	3
id02	45	9
id04	26	4
id41	10	3

TABLE 6.5: Post-processed Polyptoton

Take the following example (214), in which Rhetorica detects 'five' and 'fly' as derivational forms (cognates) of each other:

> Every night, four or **five** women will self-administer abortion pills at home and ten women will get on planes every day to **fly** abroad.
>
> *Polyptoton 214, 8th Debates statement id02*

Java (2015)'s Algorithm 2 combines WordNet's synsets and lexical relations, and affix stemming, to find a word's derivational forms. When cross-checked, WordNet[4] does not include 'fly' in the 'five' synset and vice versa. It also does not include 'party' in the 'transparency' synset and vice versa. This was the same for all of the Polyptoton occurrences which were deemed 'not cognates' in table 6.5.

Statement	Polyptoton	Polyptoton not cognates	Polyptoton affix issues
id01	42	3	3
id02	45	9	6
id04	26	4	3
id41	10	3	3

TABLE 6.6: Post-processed Polyptoton with affix issues

Cross-checking with WordNet Java (2015) indicates that the issue lies elsewhere in Algorithm 2, potentially with Java's affix removal which did not "adversely affect the algorithm results in any serious way" in the original

[4]WordNet Online: `http://wordnetweb.princeton.edu/perl/webwn`

study. However, most of the polyptoton occurrences which are 'not cognates' appear to emanate from from suffix removal (Table 6.6). Take the detected cognates transparency and party as cognates as an example (144):

> the Government must share this advice with all Oireachtas Members because we need **transparency** and informed debate above all.
>
> I agree with Deputy Billy Kelleher and my **party** colleague, Deputy Gerry Adams, that the first order of business is the repeal of the eighth amendment.
>
> *Polyptoton 144, 8th Debates statement id04*

This appears to be a stemming issue in which **'par'** is left once the prefix **'trans-'** and the suffix **'-ency'** are removed by Rhetorica. In the RF-MT use case, such polyptoton occurrences were relatively few and were detected at the human post-processing stage.

Polysyndeton

Rhetorica's search window for polysyndeton is a single sentence. For this reason, all occurrences of polysyndeton detected by Rhetorica remained after post-processing (Table 6.7). Polysyndeton is often used for rhythm and as a linguistic device, it presents a unique challenge in that it can be difficult to discern whether the overuse of conjunctions in a sentence is deliberate or natural language usage. This example of polysyndeton usage (80) detected by Rhetorica demonstrates the rhythmic nature of polysyndeton as a rhetorical feature:

> As legislators, we cannot accept the terrible impact of the eighth amendment on women's health, their obstetric care **and** well-being, **and** their **and** our fundamental rights.
>
> *Polysyndeton 80, 8th Debates statement id04*

Statement	Polysyndeton	Polysyndeton post-processed
id01	8	8
id02	5	5
id04	9	9
id41	0	0

TABLE 6.7: Post-processed Polysyndeton

On the whole, Rhetorica performs well, particularly with figures which display straightforward repetition such as epanaphora, epistrophe and polysyndeton. Rhetorica found polyptoton slightly more challenging due to the fact that the repeated word took on a different word form. It can be concluded that Rhetorica, while not perfect, is a useful tool when it comes to the automatic identification of rhetorical figures from a corpus of political speech. As with all digital tools, there is an element of 'user beware' when it comes to using Rhetorica. Therefore, combining Rhetorica's output with a human post-processing intervention is the optimum approach.

Step 4: Codify knowledge

The codification stage of this process is one of the most important, as it is the vehicle for taking tacit D.1 knowledge and representing it in a way that D.2 systems can utilise it. Through trialling different types of markup, it became clear that the complex nature of rhetorical figures makes them difficult to annotate. Take this example in which Rhetorica detects two distinct occurrences of epanaphora over the same phrases:

> The calls that have been made so far for a respectful debate on the issue have been well made. **It is** important we all recognise that

> fundamentally this is a deeply personal issue, and **it is** a matter for each individual citizen, as part of a referendum, to decide what he or she believes is right. **It is** important that all of us are non-judgmental in this regard and that we respect the views and opinions of others.
>
> *Epanaphora 64 & 65, 8th Debates statement id41*

This overlap demonstrates why inline markup is not the appropriate choice when marking up rhetorical figures. This is an issue which was also encountered by Harris et al. (2018) who, as mentioned, developed a specific annotation schema for rhetorical figure markup, the WAS-XML. Rather than using inline markup, a solution is to use standoff markup. Standoff markup stores the annotations in a separate location to the main text.

The XML format is a standard which is widely used to encode machine-readable data. As a result, it is recognised by most computer systems. This was acknowledged in the interview process:

> This is an OK format for us, we can process this one. We can actually, we use this information as a feature and whatever format it is, finally when we use it, we need to extract the relationship and represent this relationship and feature.
>
> *Round 1, Interviewee 4*

Given that XML is a widely recognised standard, and following the release of the annotation schema, the decision was made to use the WAS-XML as the markup schema for this study.

Step 5: Collaborate to evaluate

The second iteration of the interview process (outlined in Chapter 5) involved carrying out semi-structured interviews with four practitioners from the MT domain (D.2) in order to determine whether a repository of marked

up rhetorical figures could be implemented in an MT system or workflow. This was a key stage in the collaborative process. At this juncture, an interdisciplinary collaborative space was established for D.1 and D.2 researchers. In this space, expertise from both domains could not only be shared and discussed, but also be refined to create the most appropriate mode of knowledge transfer (the WAS-XML standoff markup of rhetorical devices in this case). Semi-structured interviews provided the scope to present actual examples of marked up rhetorical figures and get feedback on how these might be integrated in an MT system.

Although the research topic presents a challenge for MT practitioners, it also piqued their interest which is encouraging from the perspective of future research in this area. In terms of implementation, it remains an overall challenge to indicate to an MT system that these figures of speech are important and need to be translated consistently throughout the machine's own translation process:

> this is a huge challenge
>
> *Round 2, Interviewee 2*

Through the interview process, I was able to draw out some of the main challenges associated with integrating this type of nuance in MT systems. One hurdle to be overcome arises from the neural MT perspective. The nature of artificial neural networks is that we no longer tell the computer what to do. Instead, the neural network is trained to learn from a large set of data to produce its own solution. The result of this is that researchers have limited control over how an NMT system performs its translations.

> most people have moved on to these neural models and we call it a black box – you can't really delve in and change things as easily
>
> *Round 2, Interviewee 3*

It is not impossible to integrate marked up features in NMT systems. However, if they are integrated, it has the resulting effect that the translation becomes difficult to evaluate.

> if you integrate features marking these things, then it's also very hard to evaluate. It's hard to know what it has learned, what it has not learned because it is this black box
>
> *Round 2, Interviewee 2*

One of the challenges highlighted by Harris et al. (2018) and Dubremetz (2017) was the lack of annotated rhetorical figure data for use in machine learning tasks. This issue also emerged from the RF-MT collaborative discussions. For much work in MT, there is a lack of adequate training corpora.

> The problem is always: if you want to train an NMT system you need 2 million parallel sentences
>
> *Round 2, Interviewee 1*

The fact that these challenges exist is encouraging in terms of future work which could arise from this study.

On the whole, the D.2 experts could see value in this research into rhetorical figures, particularly from a post-editing stance.

> [I can] definitely see its importance from a post-editing point of view.
>
> *Round 2, Interviewee 2*

It is easy to envisage how a post-editing process or tool might implement this markup. A simple, yet very effective solution could be to display the source text on one side, with rhetorical figures highlighted, and the target text alongside. The human post-editor could quickly and easily see how the figures have been translated by the machine and make edits or adjustments accordingly.

> because translators are humans so they need to see that ...
> a machine can read this, but a human would spend too much time trying to look
>
> *Round 2, Interviewee 3*

Figure 6.2 depicts how a post-editing interface might present source and target rhetorical figure machine translations to a human post-editor. A human post-editor's eye could quickly be drawn to highlighted rhetorical figures detected by Rhetorica on the left, and the highlighted target language translation on the right, to ensure that repetition has been maintained consistently.

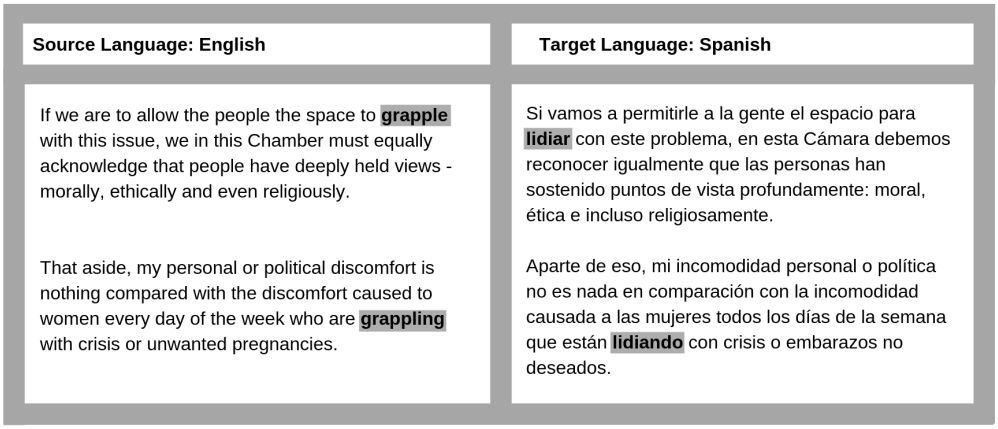

FIGURE 6.2: Mock up of how rhetorical figure highlighting in a post-editing interface might look

Ultimately, while the interviewees said they could see the value in this work, there is a feeling among them that the MT field is not at the stage where it is

considering how to implement such level of nuance in its systems. As such, they do not deal with linguistic nuances like this in the course of their work. However, this may increase in significance as the field progresses:

> I think it's definitely something that we should work on at some point, but now there are still other issues which are not solved
>
> *Round 2, Interviewee 1*

In terms of the future, there is an expectation that:

> the issues in our field will be more like these issues – it's becoming harder and harder to find little things, nuances to fix.
> ...
> It's coming I think
>
> *Round 2, Interviewee 1*

Overall, while rhetorical figures are not a current consideration for the interviewees, they could see the importance in carrying out research into integrating nuanced linguistic features in an MT workflow. An imminent implementation of this work could be in a human post-editing system to highlight rhetorical figure translations. Due to the current MT paradigm which is NMT, assimilating marked up rhetorical figures in the MT process itself appears to be further in the future. However, having seen how rapidly the field advances, that future may not be too distant.

From a collaborative perspective, semi-structured interviews provided a good environment to articulate and understand the challenges which exist in terms of marking up rhetorical figures, and integrating nuance in MT systems. The interview setting provided the scope to present examples and allowed ample time for discussion. The semi-structured nature of the interviews allowed a conversation to develop, while also lending a sense of formality to the interaction.

Step 6: Build repository

The Harvard Dataverse[5] proved to be the optimum location to create the repository and bank of exemplars associated with this study. It is open access and it is very user friendly in terms of its document upload interface. It requires very basic minimum metadata for ingestion in the repository, and the user can control any additional metadata. Finally, once the dataset has been ingested and published, it is assigned a digital object identifier (DOI) which ensures the dataset has a permanent identifier which eases location and citation, and ensures access.

6.4 Cross-disciplinary KT

In Chapter 2, challenges which accompany interdisciplinary research were outlined with specific reference to methodological and communication challenges. The relo-KT process addresses both of these challenges by inserting collaboration as a formal step in the process, and by using codified exemplars as the stimulus for this interaction.

Can relo-KT achieve cross-disciplinary KT?

The evaluation of the relo-KT process for interdisciplinary knowledge transfer presented in section 6.3 demonstrates that when the relo-KT process is enacted to transfer discipline-specific understanding from one domain (D.1) to another (D.2), each step of the process is achievable.

As shown, the relo-KT process is an iterative one, in which expert D.2 insight and suggestions which arise from the collaboration are fed back into the process in order to hone and refine the artefacts which are created (marked up rhetorical figures for persuasion in the case of the RF-MT use case).

In the RF-MT use case, cross-disciplinary KT was absolutely achieved. The collaborative interdisciplinary interactions between D.1 and D.2 shaped the interdisciplinary question and the knowledge codification. While the MT

[5]Harvard Dataverse: https://dataverse.harvard.edu/dataverse/eclarkephd

practitioners felt that marked up nuance of this type cannot be integrated in current NMT systems, there is a potential application for it in a human post-editing process.

The discussion of Rhetorica's output in section 6.3 demonstrates the requirement of disciplinary knowledge. D.1 expertise (or excellence as Huutoniemi (2010) refers to it) is needed to interpret the output data and determine what is important to retain. D.2 expertise is necessary to determine how the data can be most efficiently integrated in their workflows.

When the analytical framework for cross-disciplinary knowledge transfer was applied to the RF-MT use case (in the similar way as it was to the DH projects in Chapter 2), I found that by codifying tacit linguistic knowledge of rhetorical figures, and presenting this as marked up examples to researchers in the MT domain, cross-disciplinary KT was achieved (summarised in table 6.8).

Project	**Inter-disciplinary**	**Knowledge Codification**	**Evaluation**	**Cross-disciplinary KT**
RF-MT	linguistics MT	WAS-XML	semi structured interviews	iterative process; articulated challenges; understood challenges; scope to discuss; scope to present examples

TABLE 6.8: Analytical framework for cross-disciplinary knowledge transfer applied to the RF-MT use case

6.5 Summary

Overall, the relo-KT process proves effective in the transfer of knowledge across disciplines. Its efficacy depends on a multi-step, iterative process which relies on presenting codified tacit D.1 knowledge to D.2 practitioners in a collaborative scenario.

The development of the relo-KT process was motivated by the gap which emerged from the literature which demonstrated that despite being cross-disciplinary and collaborating to produce interdisciplinary research, many research project teams do not document their collaborative approaches. The work of Szostak (2002) and the PMI (2015) were instrumental in developing the steps of the process and the framework for analysing cross-disciplinary KT in DH projects.

The RF-MT use case which was used to trial the relo-KT process in an interdisciplinary setting arose from a gap which emerged from literature on rhetorical figures and their role in MT systems. This gap is potentially due to the fast-paced evolution of the MT domain.

Chapter 7

Conclusion

> Science and art sometimes can touch one another, like two pieces of the jigsaw puzzle which is our human life, and that contact may be made across the borderline between the two respective domains.
>
> *M. C. Escher*

7.1 Thesis summary

At the beginning of this thesis, I set out to examine whether a method for cross-disciplinary knowledge transfer could be formalised and realised. Through the preceding chapters, I have developed, tested and presented a potential method for cross-disciplinary knowledge transfer which can be enacted in an interdisciplinary scenario. In this final chapter, I will review and summarise the work which has been presented through the preceding six chapters and I will address the research questions and objectives. Before concluding, I will outline the contributions this thesis makes while also addressing the limitations of the study. Finally, I will outline some suggestions for further research in this area.

7.2 Review of preceding chapters

Chapter 1 presented the background and motivation for this work. It also introduced a potential gap in the interdisciplinary sphere as highlighted by

Griffin and Hayler (2018) in their 'Collaboration in Digital Humanities Research – Persisting Silences' paper. The gap relates to the silence when it comes to discussing interdisciplinary collaboration. This suggested that there is room for a formalised method of cross-disciplinary knowledge transfer.

To address research objective 1 (RO1), Chapter 2 presented the literature relating to interdisciplinary research and knowledge transfer processes. The chapter started by defining key terminology including Choi and Pak (2006) who define interdisciplinary research as research which "analyses, synthesises and harmonises links between disciplines into a coordinated and coherent whole". Choi and Pak's definition was referred to multiple times as this thesis progressed. Two other key pieces of work which were referenced throughout this thesis were Szostak (2002)'s 'How to do interdisciplinarity: Integrating the debate' in which he outlines twelve steps for doing interdisciplinary research and the PMI (2015)'s knowledge transfer life cycle. These were integral in developing both the relo-KT process and the analytical framework used to evaluate digital humanities projects' achievement of cross-disciplinary knowledge transfer.

Further to this, Chapter 2 also included analysis of five DH projects. By applying an analytical framework for cross-disciplinary KT, it was deduced that while the collaborative approaches employed by each project team were not explicitly documented, they clearly occurred. This finding exposed a gap in terms of formalising cross-disciplinary knowledge transfer which the relo-KT process was developed to fill.

Chapter 3 fulfils RO2 in that it outlined the relo-KT process, a formalised process for cross-disciplinary knowledge transfer which draws on the literature and interdisciplinary challenges presented in Chapter 2. The relo-KT process consists of six main steps which take an interdisciplinary approach through the iterative, collaborative process. The process begins with a collaborative effort to establish an interdisciplinary question, before moving on to the data acquisition, extraction and codification step. Further collaboration occurs to evaluate and improve the knowledge codification before a satisfactory artefact is finalised. The collaborative, interactive approach is a key feature of the relo-KT process.

In Chapter 4 additional background was presented, to further address RO1. Literature on rhetorical figures, and the linguistic theory related to them was reviewed. So too was the theory and background of MT. Linguistics and MT are the two domains which this work brings together, to transfer understanding between them. The first half of the chapter dealt with persuasive language, namely rhetorical figures and how they are used in political speech. Key work in the area of computational rhetoric was presented including research on the automatic detection of rhetorical figures by Gawryjolek (2009), Harris and Di Marco (2017), Dubremetz (2017), and Java (2015) who developed the Rhetorica software used to detect rhetorical figures from a corpus. Rhetorica can detect 14 rhetorical figures in total. In this work I focused on four of them: epanaphora, epistrophe, polyptoton and polysyndeton.

The second half of Chapter 4 presented a history of Machine Translation (MT) from its beginnings in Rule based MT (RBMT) to Example based MT (EBMT), Statistical machine translation (SMT) and recent developments in Neural machine translation (NMT). The MT and rhetorical figures section (4.8) highlighted a gap in the related literature. The limited work around rhetorical figures and translation has mainly been carried out in the context of advertising and transcreation, in which translators produce a translation which is similar to the original, but can also evoke a reaction in the target language (Shriver, 2011). Smith (2006)'s work on comparing translations of rhetorical figures in English and Russian magazine advertisements demonstrates this nicely. The research which has been carried in the area of rhetorical figures and translation has taken place in the human translation domain and it is something which to date has not been focused on from an MT perspective.

Chapter 5 presented a use case to fulfil RO3. The RF-MT use case applied the relo-KT process to the domains of linguistics (rhetorical figures) and MT. In order to achieve KT across these domains, a corpus of political speech was created from statements delivered in Dáil Éireann. Java (2015)'s Rhetorica software was used to extract occurrences of four rhetorical figures from the corpus. From Rhetorica's output, linguistic understanding of exemplar rhetorical figures for persuasion was codified. This codified knowledge was presented to NLG and MT experts in an iterative cycle of semi-structured interviews and their insight and suggestions influenced further iterations of the process.

Finally, Chapter 6 reflected on each distinct piece of the puzzle (the relo-KT process, the RF-MT use case and the analytical framework) and combine them in order to assess the overall method and approach taken. The evaluation of the relo-KT process demonstrated that a) the process as a whole is implementable and b) is successful at transferring tacit knowledge from one discipline to another. The success of the process hinges on the 'knowledge codification-collaboration' iterative loop which ensures that recommendations and direction which emerge during the interactive process are captured and fed back into the process.

7.3 Research questions

This study aimed to answer two research questions:

RQ 1: How can a method of cross-disciplinary knowledge transfer be formalised (e.g. transferring linguistic understanding to MT)?

RQ 2: To what extent can such a formalised method support the identification of persuasive rhetorical figures in a corpus; and to transfer linguistic understanding of them, to support Machine Translation (MT)?

By applying the relo-KT process in an interdisciplinary use case (the RF-MT use case), this study found that the process can be sucessfully implemented to transfer tacit knowledge between disciplines. The RF-MT use case transferred codified linguistic understanding of persuasive rhetorical figures to the MT domain through an iterative series of semi-structured interviews.

7.4 Research contributions

In the course of this thesis, a method for cross-disciplinary knowledge transfer was formalised and represented as the multi-step relo-KT process, which can be clearly understood and followed.

The process was enacted in the RF-MT use case after Chapter 4 demonstrated that shortfalls remain in MT systems particularly when it comes to integrating the nuances of natural human language such as rhetorical figures.

7.4.1 The relo-KT process

The relo-KT process was developed as a set of steps which can be followed in order to achieve cross-disciplinary knowledge transfer. This process can be used to formalise collaborative efforts in interdisciplinary research. It can also complement existing frameworks such as Szostak (2002)'s twelve-step process for interdisciplinarity which influenced the development of this process.

Each step of the relo-KT process is implementable, from the establishment of an interdisciplinary question to the acquisition of a corpus of data suitable for analysis and the extraction of domain specific understanding from it. The nub of the process involves knowledge codification and collaboration to convey understanding across disciplinary boundaries. Tacit knowledge transfer develops between the discipline experts as the process iterates.

The overall aim of developing the relo-KT process was to formalise the knowledge transfer interaction between disciplines. The successful application of the process in the RF-MT use case indicates that it achieves its aim. When the analytical framework developed in Chapter 2 was applied to the RF-MT use case, it demonstrated that the relo-KT process could be used to document cross-disciplinary knowledge transfer. The formalisation of the interdisciplinary exchange ensured that the scope existed to articulate challenges, present exemplars and discuss solutions. The use of semi-structured interviews ensured the collaborative interaction was documented.

7.4.2 Secondary contributions

RF-MT

A secondary, but nonetheless important focus of this study was the implementation of the relo-KT process to identify rhetorical figures in a corpus of political speech to demonstrate how the linguistic understanding of such devices might be transmitted and integrated in an MT workflow. Until now, there has been little research carried out on integrating rhetorical figures in MT workflows. Findings which emerged from interviews with MT researchers indicate that the nuance of rhetorical figures could be implemented in a human post-editing workflow, but this is something which is not a priority for those working with NMT systems currently, although this may well change quite quickly as MT tools continue to become more sophisticated.

Corpus and exemplars

RO4 was to assemble a repository which contains a corpus of political speech and a bank of rhetorical figures. This is a less significant contribution, but nonetheless is beneficial, should someone want to reproduce or advance the research which was carried out in the course of this thesis. This is accessible at the following location:
`https://dataverse.harvard.edu/dataverse/eclarkephd`

7.5 Research limitations

The relo-KT process is a time consuming one. The process requires collaborative iterations which involve talking to researchers with specific expertise and qualitatively assessing their input at various stages of the process. Subsequently, it involves producing and refining artefacts based on the feedback and insight received in the interviews. Therefore, it is necessary to allow ample time for the process to iterate and evolve.

The interviews were semi-structured in nature which allowed conversation to develop organically. The organic nature of the interviews meant that as each interview evolved, the conversation branched in different directions. As a result, the exact prepared questions were not asked in all interviews. This led to a minor challenge when it came to generalising the interview data. To address this in future implementations of the process, a follow-up questionnaire with the interview participants could complement the qualitative data.

There is always a limitation when using a tool developed for a purpose other than the study at hand. In this work, every effort has been made to ensure that Rhetorica had been developed methodically. When anomalies arose (such as affix removal in polyptoton (section 6.3)), they were approached critically and reported on transparently.

Corpus size depends on the specific linguistic research query. For big data analytical approaches, a large corpus is a requirement. However, with large corpora come drawbacks such as noisy data. While the specialised 8th Debates corpus was a sufficient size for the exploratory nature of this research, another research question may require a different approach to corpus building. For example, to measure rhetorical figure usage over time, a diachronic corpus might be required.

7.6 Future work

The relo-KT process could be applied to different domains which require interdisciplinary interaction to transfer tacit knowledge or expertise in a meaningful and measurable way. A future application of the relo-KT process could be to develop further understanding for dialogue systems. Also known as conversational interfaces or chatbots, dialogue systems are computer systems which aim to converse naturally with a human. Conversational interfaces present a challenge to move away from rigid, robotic interactions towards replicating authentic human speech style. Understanding the elements of human speech is clearly an interdisciplinary task, which can involve combining expertise from the speech synthesis discipline with phonetics and/or acoustics expertise, for example. If the emphasis is on the very

complex task of creating a conversational interface which speaks with a particular accent, the relo-KT process could be applied to formalise knowledge transfer between a sociolinguist, who has expertise in accent and dialect, and a speech synthesis expert.

The manual mark up and close reading approaches required to become familiar with the corpus are time-consuming. The findings gleaned from the human post-processing aspect of this thesis could be used to develop an automated markup system for rhetorical figures. This would remove some of the human outlay at this stage.

An opportunity also arises to integrate rhetorical figure understanding in a post-editing system for MT. A potential application in a post-editing environment was visualised in Chapter 6. In a nutshell, a post-editing interface might present source and target rhetorical figure machine translations to a human post-editor. The human post-editor's eye could be drawn to highlighted rhetorical figures detected by Rhetorica on one side, while the target language translation would be highlighted on the other, to ensure that repetition has been maintained consistently.

Finally, the RF-MT use case discussed in this work focused on four of the potential fourteen rhetorical figures that Rhetorica can identify. Future work could implement Java (2015)'s software to explore different rhetorical figures. For example, the use of the use of paradox in a literary corpus could be explored by examining Rhetorica's detection of oxymoron.

Appendices

Appendix A

The 8th Debates corpus

The 32nd sitting of Dáil Éireann commenced on 10 March 2016 with 158 elected members (TDs). The gender balance is 35 female TDs to 123 male. One of the most contentious issues to face the 32nd sitting of Dáil Éireann in the first two years of its five year sitting was that of the 8th Amendment. The 8th Amendment[1] which recognised the equal right to life of a pregnant woman and the unborn saw the insertion of a new Article 40.3.3 into the Constitution of Ireland following a referendum in September 1983. In the intervening three decades, there were a number of constiutional and legislative changes relating to Article 40.3.3, notably the insertion of the 14th Amendment, the right to travel to another jurisdiction for a termination of pregnancy following the X Case in 1992, in which a 14-tear-old became pregnant as a result of rape. A High Court ruling prevented the girl from travelling for an abortion. This was later overturned following an appeal to the Supreme Court and permission to travel was granted.

It was another 26 years before the opportunity to repeal the 8th Amendment was brought before the Irish people. In 2016, the then Taoiseach, Enda Kenny, establised The Citizen's Assembly which was described as "an exercise in deliberative democracy, placing the citizen at the heart of important legal and policy issues facing Irish society"[2]. The Assembly comprises a chairperson and 99 citizens, "randomly selected to be broadly representative of the Irish electorate". The Assembly met on five weekends between 26 November 2016 and 23 April 2017 to consider the 8th Amendment and made recommendations to the Oireachtas which included that "Article 40.3.3 be replaced with a Constitutional provision explicitly authorising the Oireachtas to address termination of pregnancy, any rights of the unborn and any rights of the pregnant woman" (*The Eighth Amendment of the Constitution - The Citizens' Assembly*).

A Joint Committee of both Houses of the Oireachtas - Dáil Éireann and Seanad Éireann - was established to review the recommendations of the Assembly. The Committee met in public session between September and December 2017 and published a report on 20 December 2017 which concluded that "we need some change and in order to effect that we need to

[1] Article 40.3.3 wording: The State acknowledges the right to life of the unborn and, with due regard to the equal right to life of the mother, guarantees in its laws to respect, and, as far as practicable, by its laws to defend and vindicate that right. `http://www.irishstatutebook.ie/eli/cons/en`

[2] The Citizen's Assembly: `https://www.citizensassembly.ie/en/Home/`

amend the Constitution to remove article 40.3.3. After many years of public and political debate on the issue, the people will have their say" (Oireachtas, 2017).

Over five dates in January 2018, TDs in Dáil Éireann were permitted to give statements on the Report of the Joint Committee on the 8th Amendment of the Constitution. When the statements were delivered in Dáil Éireann, the referendum was being planned which would put the question of repealing the amendment to the people, effectively reducing it to a Yes / No question. The TDs who declared on which side of the divide they fell were advocating for one side or the other. For this reason, persuasive elements are sure to have be woven into any statement they made on the subject prior to the referendum.

Of 158 sitting TDs elected to the 32nd Dáil Éireann, 63 made statements in response to the Report. Of the 63 TDs who made speeches, 45 were in favour of repealing the 8th Amendment while 16 were against. Two TDs who spoke on the issue had not declared their position at the time.

The 8th Amendment was a divisive subject given that the moral and ethical concerns around the issue of abortion and reproductive rights. A number of political parties including Fine Gael (lead party) and Fianna Fáil (main opposition party) allowed their members a conscience vote on the matter. Given that the debate topic was emotive, controversial and polarised and that the 8th Amendment had a powerful legacy since its insertion in the Irish Constitution in 1983, it can be assumed that any TD who chose to speak on the record about the topic would have crafted their statement carefully in order to avoid misinterpretation or obfuscation.

The referendum to repeal the 8th Amendment took place on 25 May 2018. Throughout the referendum campaign, the Irish Times tracked TDs declarations in favour of or against repealing the amendment and categorised these into three groupings. In response to the following question: "Should the 8th Amendment to the Constitution be repealed?", 'Yes' indicated that the TD is in favour of repeal, 'No' denoted that they are against repeal and wished to keep the 8th Amendment in the constitution with no change to its wording. The final category was what the Irish Times have termed 'Undeclared'. For

the purpose of this case study, the Irish Times terminology and tracker results are used to classify whether individual TDs were for or against repealing the 8th Amendment ("Referendum Tracker").

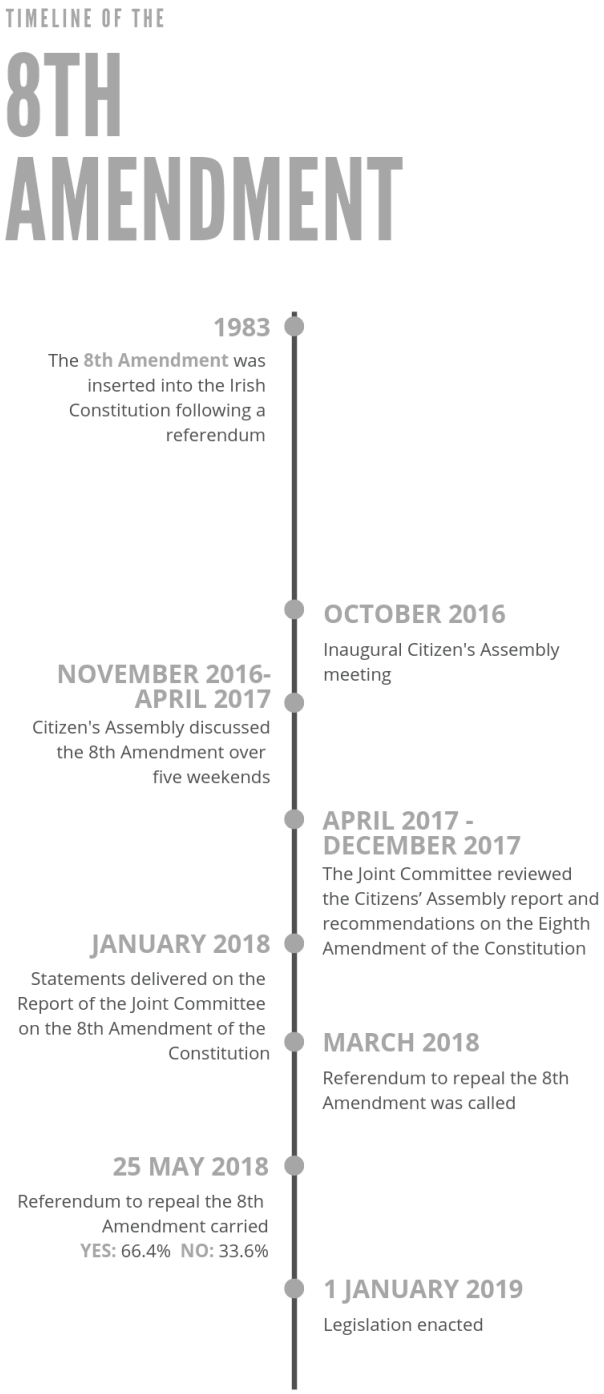

FIGURE A.1: Timeline of significant events related to the 8th Amendment of the Constitution of Ireland

Appendix B

Sample TD speeches from the 8th Debates corpus

Record id: 1
TD name: Simon Harris
Date: 17 January 2018

Every now and then an issue comes before us which challenges us to think about what kind of a country we want to be and what kind of a society we are; an issue that we struggle with, that may be difficult to talk about, but that is not going to go away. Today is a moment where we, the Members of the 32nd Dáil, come face to face again with such an issue. In doing so, we come face to face with our history - a history that continues to unfold and continues to hold up a mirror in which we sometimes do not like what we see, whether it is the damp cold of the Magdalen laundries creeping into our bones, or the sundered silence of mother and baby homes being broken, or the glimpses of what was an all too acceptable culture exposed by the Kerry babies case. All these are connected by the way we as a country have treated women, and particularly the way we have treated pregnant women.

I think of another cold January like this one in 1984 when the 15 year-old Ann Lovett gave birth alone to her son beneath a statue of Our Lady. The death of Ann and her baby son in these stark and lonely circumstances is a memory that chills us still and one we should not forget. We now arrive at another moment on a long journey, starting with the insertion of the eighth amendment of the Constitution in 1983 through the court cases that made us think about the pregnant victims of rape and incest, through the bravery of families faced with fatal foetal abnormalities who made us think about the particular cruelties we add to their tragedies, and, after all this, perhaps arriving at the realisation that each crisis pregnancy is different and each involves a real woman facing a very difficult and very personal decision.

These are real women such as the 36 from County Carlow who travelled to the UK for an abortion in 2016, or the 38 from Mayo, the 69 from Tipperary, the 85 from Wicklow, the 241 from Cork and the 1,175 women from Dublin. Women from every county in the Republic travelled to the UK in 2016 and we need to acknowledge them all, including the 49 from Kerry; 130 from Kildare; 21 from Leitrim; 20 from Roscommon; 69 from Wexford; 39 from Cavan; 15 from Monaghan; 99 from Limerick; 53 from Clare; 38 from Westmeath; 63 from Donegal; 113 from Galway; 44 from Kilkenny; 42 from Laois; 83 from Louth; 100 from Meath; 28 from Offaly; 29 from Sligo; 16 from Longford; and 56 from Waterford.

In 2016, 3,265 Irish women travelled to the UK alone and we know that Irish

women travel to other countries such as the Netherlands as well. More than 1,200 of the women who went to the UK were aged between 30 and 39; more than 1,500 were aged between 20 and 29; 255 were aged 40 or over; ten were girls under the age of 16; and 230 were teenagers. More than half of the women who travelled were married, in a civil partnership, or in a relationship while 85% of them were between three and 12 weeks' pregnant. It is estimated that at least 170,000 Irish women have travelled to other countries for abortions since 1980.

These are not faceless women. It might be convenient for us sometimes to think that they are. They are our friends, neighbours, sisters, cousins, mothers, aunts, and wives. Each woman is dealing with her own personal situation and making what is a deeply difficult decision because this time around - let us be honest about this – this is not a decision or a procedure that anyone undertakes lightly. Women agonise about it and consider every possibility for dealing with the particular crisis facing them, and sometimes they arrive at the conclusion that there is no other option for them but to terminate their pregnancy. When they arrive at that difficult decision, the country we live in, which we hope has come a long way from the dark events that continue to haunt this Chamber, tells them to go and get their care elsewhere - go to another country or head off somewhere else.

In 1992, we formalised the right of Irish women to travel for an abortion and to obtain information about it, but we have been temporarily exporting women in crisis for an awful lot longer than that. I cannot help but wonder what we would have done if we did not have a neighbouring island to help us turn a blind eye. Sometimes turning a blind eye is the same as turning your back. We need now to seek to build a society which accepts our own challenges and addresses them honestly, maturely and openly, which does not seek to deny reality or to outsource it to another country, and which does not reject women at the most vulnerable moments in their lives.

As I stand before this House at the commencement of what I genuinely believe in time could be seen as an historic debate, I am fully aware of the sensitivities and complexities of this issue. I want to acknowledge the deeply held, genuine views on all sides of the House and throughout the country. No matter what may divide us, I accept that all of us are trying to do what is right. All of us are guided by our own conscience and our own sense of humanity. Some of us have changed our views over the years. My own views have changed and been formed by listening to women and doctors, and coming to recognise some hard realities. Some of us bear the scars of past debates

and fear what is to come. However, this time, I firmly believe that it is possible for us to have a respectful debate on the issue. Please do not call that naïve and do not dismiss the idea that we can maturely recognise that each of us has deeply personal and genuinely held views, all of which deserve to be heard, understood and respected. It is an issue that troubles most of us as individuals. For some of us, it challenges us to hold what appear to be conflicting views simultaneously. Which of us does not value and love human life, and which of us does not want to see that protected? No one has a monopoly on that. The tactics name-calling, pigeon-holing, and stereotyping need to be consigned to history because they have only led to paralysis, fear and division.

It will require effort and attention from all of us, regardless of our views, but it is so important that everyone has the chance to hear clearly in order that when, as a nation, we come to make the next decision on this issue, it is informed.

We do so as a country with a particularly complex past which, in fact, dates back to 1861, when abortion was a felony under the Offences against the Person Act, a felony with a sentence of penal servitude for life. In more recent decades, it has been an issue dominated by referendums and court cases. As Members know, 1983 saw the first referendum. In 1992 there was another with three questions. Legislation followed in 1995. A third referendum on abortion was held in 2002 seeking to overturn the X case but it was defeated. In 2014 the issue came before this House again when we passed the Protection of Life During Pregnancy Act.

I remember vividly that debate and some of the offensive comments about floodgates opening. I remember the language used, which seemed to suggest women would even fake a threat to their own lives to obtain a termination - quite unbelievable really, when we look back only those few short years. Obviously, none of this has come to pass and the reports laid before this House each year bear that out.

Since the passing of that law there has been a clear legal basis for abortion in Ireland but it has become clear that the Oireachtas can go no further without constitutional change. Other Members have tried to bring forward thoughtful legislation to assist families with a diagnosis of fatal foetal abnormality, for example. I have been the Minister to respond to these Private Members' Bills but on each occasion the legal advice has been clear that without the repeal of the eighth amendment, we, as an Oireachtas, could not address these issues.

Abortion is a reality for women living in Ireland but not just women in the limited circumstances in which it is legal under the Protection of Life During Pregnancy Act, nor for the many women who travel to other countries, as I have outlined. There are now new realities on top of that. The Oireachtas committee heard evidence of abortion pills being bought on the Internet and used by women in this country without any medical supervision. Research from the British Journal of Obstetrics and Gynaecology shows a 62% increase in the number of women from Ireland contacting one online provider over a five-year period, up from 548 women in 2010 to 1,438 in 2015, and that is just one provider.

Can we pause and picture what this is telling us, because we can get lost in numbers and years? Is it acceptable to any of us that women are once again left in a lonely and scary place, sending off for a pill to be sent through the post instead of being able to access the medical advice and support they need? This is happening in Ireland today. It is a fact. How can we ignore it? How can we consider it to be all right? If it is the sad reality that we have been exporting this issue for many decades, are we now accepting that, on top of exporting it, women must import their own solutions?

I want to turn now to the substance of the recommendations of the Joint Committee on the Eighth Amendment of the Constitution. I commend all the members for their work and thank them for the contributions they made. They have served the Oireachtas well. I wish to thank Senator Catherine Noone, in particular, for her calm and balanced handling of the issue as Chair. We, as an Oireachtas, asked these colleagues, on a cross-party basis, to do a very important body of work: to listen to experts, to hear evidence and to report back to us. We owe them a debt of gratitude for the time and dedication they applied to their task. I would also like to commend the chair of the Citizens' Assembly, Ms Justice Mary Laffoy, and its members for their careful deliberations and to acknowledge their valuable contribution.

The Citizens' Assembly and the committee have given us a model for addressing this issue in a rational and measured way, and I believe it is one we should follow. I want to recognise that the recommendations contained in the committee's report represent the views of the majority of members but there was not unanimous agreement on them. I respect the views of those who dissented from the recommendations but I believe the recommendations are the basis on which we must proceed on this issue. The main conclusion of the committee's work is that change is needed to extend the grounds for lawful termination of pregnancy in the State. In order to effect that change, the

committee recommended that Article 40.3.3o should be removed from the Constitution. The committee then went on to make recommendations on the grounds on which termination of pregnancy should be permitted in Ireland, if Article 40.3.3o is repealed by the Irish people. It recommended extending the law on abortion to cover cases where the health of a woman is concerned, cases of fatal foetal abnormalities and a broader legal regime that allows abortions where the woman seeks it from her medical practitioner if her pregnancy is under 12 weeks gestation.

I am working with my chief medical officer and officials of my Department, and the Attorney General, to consider how best to translate these recommendations into legislation, should that be the wish of the Irish people. It is my intention that, in the event of a referendum, as much information as possible would be available to people so they can make an informed decision.

While it is understandable the focus so far has been on the committee's recommendations regarding the eighth amendment, it is important to put on the record of the House that the committee did not only make recommendations on termination of pregnancy, but also on the services and supports that should be available to women. I am fully committed to ensuring that all women accessing maternity services in our country should receive the same standard of safe, high quality care. Every woman, from any corner of Ireland, should expect and be able to access the maternity services she needs. I am confident that, through the implementation of the first-ever national maternity strategy, Creating a Better Future Together, the quality outcomes envisaged by the committee will be realised. In some ways, it is incredible it is the first-ever national maternity strategy. Officials in my Department, under the chairmanship of the chief medical officer, have now established a group to address and formulate an effective and comprehensive response to the issues raised by the committee in its ancillary recommendations.

We have made other progress which provides the base for delivering the kind of integrated care women and their babies deserve. We have established the national women and infants health programme. We now have HIQA's standards for safer better maternity services and new HSE national standards for bereavement care to ensure clinical and counselling services are in place to support all women and families in all pregnancy loss situations. The HSE's Positive Options crisis pregnancy counselling service is also available in 50 centres nationwide.

As someone born three years after the 1983 referendum on the eighth amendment, I never imagined I might one day be the Minister for Health responsible for a referendum on its repeal. I come at it from a perspective that I think was sadly absent in 1983, that is, from the perspective of women's health care. In the Ministry I have the honour to hold, it is my duty to work to ensure that people in this country receive the highest possible standards of care, and to protect and promote the health of our people under the laws of our land.

I realise the issue before us challenges us - it challenges me. It causes us to ask difficult questions of ourselves. It makes us uncomfortable at times as we collectively wrestle with what is, at its core, a very personal and private matter. Women become pregnant and it is a joyous thing for so many, but it is a terrifying thing for some and a tragic thing for others. Irish women are driven today to find their own solutions. Sometimes they put themselves at risk in doing so. As things stand, they are often left without help, advice or support at one of the most vulnerable times in their lives. I hope that, as a country, we can no longer tolerate a law which denies care and understanding to women who are our friends, our sisters, our mothers, our daughters, our wives. Ultimately, there is always a deeply personal, private story behind each individual case, which I believe is a matter best served by women and their doctors. I believe the people in this country trust women and trust their doctors to make these difficult decisions.

I look forward to what I hope will be a constructive debate on the issues raised by the committee here and in the Seanad. I hope we can show here that this debate can take place in an atmosphere of respect for each other's views so that the same is possible in the context of a referendum campaign. After this debate concludes, I expect to return to Government in the coming weeks with a series of proposals which I believe can deliver a referendum by the end of May or very early June, should the will of the Oireachtas be to facilitate that. I do not doubt that, as long as I remain a Member of this House, I will continue to witness moments in this Chamber that remind us of darker times in our history, but let this be a different type of moment. Let this be a moment people will look back on as one where their representatives confronted one of the most complex issues we have faced as a country with clarity, with compassion and with care.

Record id: 2
TD name: Billy Kelleher
Date: 17 January 2018

I welcome the opportunity to speak on this divisive issue, which has been at the heart of debate since 1983 and the insertion of Article 40.3.3° into the Constitution. Like the Minister, I was too young to vote in that campaign, although I am a few years older than him.
I was only three years short of being able to vote in 1983. All of this indicates that there is a large swathe of people who have never had an opportunity to cast an opinion on this issue. It should be borne in mind that it is primarily women of childbearing age who are most affected, yet they have never had an opportunity to cast their view.
Since being appointed Fianna Fáil's spokesperson on health, I have been grappling with this issue. We debated it in the context of the Protection of Life During Pregnancy Act. I sat on the committee chaired by now Senator Buttimer that formulated that debate. It was insightful because that was the first time that we as an Oireachtas tried to deal with the issue in a non-political way, recognising that it was socially divisive and people had strong feelings on it.
My leader, Deputy Micheál Martin, subsequently allowed us a free vote - a vote of conscience. Some parties have followed us in that while others have not. That is entirely their entitlement. It is not for me to say whether they should or should not. Since being appointed health spokesperson, I have tried to take the political element out of this issue, given that it has been charged for a long time and has occasionally been used by political parties of all persuasions and none. We have gone beyond that now. Our society is mature enough to have a respectful, incisive and decisive debate on this issue, with the people ultimately deciding.
I also sat on the joint Oireachtas committee. I pay tribute to the Citizens' Assembly, chaired by Ms Justice Mary Laffoy, who outlined the assembly's workings and the reasoning behind its conclusions and recommendations. Our committee was charged by the Oireachtas to examine those recommendations and reach our own determinations.
The committee did not reach a unanimous view. Rather, there was a majority view and a dissenting view, with varying views even within those. I respect every one of those views. If we are to allow the people the space to grapple with this issue, we in this Chamber must equally acknowledge that people

have deeply held views - morally, ethically and even religiously.
That aside, my personal or political discomfort is nothing compared with the discomfort caused to women every day of the week who are grappling with crisis or unwanted pregnancies or the devastating news of fatal foetal abnormalities. We must be conscious of that.
The Minister outlined that 170,000 plus women had left this State for terminations. We have been exporting our problem for a long time. It is our duty to address this issue. With the best will in the world, and although people have varying opinions on this, we cannot do anything other than what we have legislated for already unless we repeal, amend or replace Article 40.3.3°. We have to change what is in our Constitution. That will be the first step towards addressing this issue.
It would be easy for me and others contributing in this debate to keep our heads down and hope that the issue goes away but, as a generation of politicians, we have to deal with it. We have to give the people an opportunity to express their opinion in light of the fact that the last time they had such an opportunity was in 1983. Ireland has fundamentally changed in many respects since then, as outlined by the Minister. It has changed most in how women have become more assertive. They now have an opportunity to put themselves at the heart of this debate, which is primarily about women's rights and health care and about giving them equal opportunity in the Republic. We cannot have a situation in which a woman becomes a second-class citizen upon becoming pregnant. We must understand that that is no longer acceptable.
People have asked me how the committee arrived at our recommendations. The Citizens' Assembly made 13 recommendations, the most fundamental one being that there should be a change to Article 40.3.3°. The assembly outlined three options. The committee's majority recommendation was for a repeal simpliciter, that is, the article should just be removed from the Constitution, and for the Houses of the Oireachtas to be allowed to legislate for what was primarily a health care issue for women. There were varying views on that in the legal advice that the committee sought and was given. I am sure that different advice may even be given to the Government by the Attorney General. What I do know, however, is that if we continue allowing Article 40.3.3° to be the bulwark for dealing with this issue, the status quowill prevail. Every night, four or five women will self-administer abortion pills at home and ten women will get on planes every day to fly abroad. We as a Parliament will have to live with that if we fail to grasp this issue.

We must show political leadership by having a debate that is respectful of every view that will be expressed in this Chamber and then allowing the people to make their decision based on the full information. I urge everyone involved not only to respect one another's views, but to engage in the debate so that the question can be decided one way or the other.

I have expressed my view that there must be a repeal simpliciteror a variation of same that removes Article 40.3.3° from the Constitution. I am open to a new constitutional article giving the Oireachtas the supreme authority to legislate on this issue. I would take on board the Attorney General's opinion and that of others. However, we must be honest with people. If Article 40.3.3° is replaced or repealed, the Legislature will govern from then on, but we cannot guarantee anything. We could publish legislation in advance of this referendum campaign that would give people guidance as to what the likely outcome would be, but if the Oireachtas is to legislate in the event of a repeal, then it will be the Oireachtas that will decide, not necessarily this one either, but possibly a changed Oireachtas following a general election. All we can do is be honest with people about what we believe the Oireachtas in its current composition would pass in the event of a repeal.

I urge people to examine the committee's recommendations and consider why we came to those determinations. They are more conservative than the recommendations of the Citizens' Assembly. For example, the assembly referred to a 22 week gestational period whereas we are calling for a 12 week period in the majority of cases. It would be a matter for the clinicians and the woman with a view to her health and life.

Many people have stated their desire to address the issue of fatal foetal abnormalities. I have a friend whose wife received a diagnosis of a fatal foetal abnormality. They wanted to have this child. They kicked the sand up and down Garryvoe beach for weeks on end because of that diagnosis wondering what they would do. Would she continue with the pregnancy or would she have it terminated? They spent weeks deciding. If they decided to terminate the pregnancy, given that there could have been no survival outside of the womb, the only difficulty would have been that they would have had to go abroad. They made the decision to go abroad. It was inhumane that our country could force this on our citizens. They decided to terminate the pregnancy because they felt that they could not go through with it in light of the prognosis of certain death ex utero, so the idea that we would ask people like them to FedEx their little baby back home to Ireland is inhumane. For that reason, the recommendations in the report are critical to this debate.

The issue of rape and incest is one about which a lot of people speak. Even people with quite conservative views in this area will say the issue of incest and rape should be addressed. All of the empirical legal evidence we have shows that it would be impossible for us to legislate on the ground of rape. We have to trust the integrity and honesty of women. Are we to put people who have been subjected to a vile act through another inquisition to find out and determine whether they have been raped in advance of having a termination? If we are to have restrictions in this area, these are the things that will actually have to happen. That is why, in the context of the recommendations of the committee, we also looked at the 12-week limit and decided that it should be a matter for the woman and her clinician to determine the reasons for a termination within that period. That would also address the issue of incest and rape in the vast majority of cases in that we would not need to have an inquisition of women to establish whether they were telling the truth. I firmly believe we should trust women in this context and that we have a duty to ensure they have the space, with their clinicians, to make the best decision for themselves and the lives they are living.

The Minister spoke about the age profile of those travelling abroad for a termination. I have received several emails from people which, to say the least, I find quite alarming. I am alarmed that there are people in our society who think this way because in their emails they are effectively saying the women who travel abroad are just using abortion as another form of contraception. I find that grossly offensive. I do not know of any woman who will get onto an aeroplane lightly and head to Liverpool, Birmingham or Holland for a termination of a pregnancy. I know women who have had terminations, an issue with which they grappled before they made their final decision. That it is just another form of contraception is a grotesque, offensive view of women.

As my party has allowed members a free vote, I am speaking in a personal capacity. I am speaking as a Fianna Fáil Deputy but in a personal capacity and as a representative of the people of Cork North Central. I leave my personal and political views outside the door when I speak about these issues because I have listened to the evidence given at the committee, heard the personal testimonies about what people had to endure and spoken to individual constituents who brought to my attention the challenges they faced when they received the devastating news of a fatal foetal abnormality or, in some cases, a foetal abnormality and the decisions they had to make. Space should be given in this debate. If we expect the people to behave in a manner that allows respect for varying views, the very least we should do is expect

the same from each other in this House. If we are to lead by example in any way, we must allow space for varying views to be expressed.
Committee members held varying views on the issues before it, but majority views were expressed in a number of areas. The Citizens' Assembly made recommendations to allow a termination for various reasons, the first of which was a real and substantial physical risk to the life of the woman. Reason No. 2 was a real and substantial risk to the life of the woman by suicide. Reason No. 3 was a serious risk to the health of the woman. Reason No. 4 was a serious risk to the mental health of the woman. Reason No. 5 was a serious risk to the health of the woman. Reason No. 6 was a risk to the physical health of the woman. Reason No. 7 was a risk to the mental health of the woman. Reason No. 8 was a risk to the health of the woman. Reason No. 9 was pregnancy as a result of rape.
As a committee, we concurred, by and large, with the recommendations of the Citizens' Assembly. However, we did deviate from some of its recommendations, primarily in the area of disability. We were very clear that disability should not be a ground for a termination. Let us be honest - people make decisions and choices. Are we to adjudicate on and be judgmental about the decisions they make? Who are we to judge a person who has just received the most devastating news that the baby she is carrying has a profound disability or will die ex-utero? The committee did state, however, that after a period of 12 weeks, disability would not be a ground for a legal termination. That was the strongly held view of a lot of committee members, even those who had quite open views on this issue. It was not the case that the committee was a slave to the Citizens' Assembly. It deliberated on every one of the recommendations and came to its own conclusions and did diverge in certain areas.
I fervently hope the Government can keep to its timetable and that we can have the substantive issue of the repeal of Article 40.3.3odealt with by the summer. Publishing the scheme of a Bill in that timeframe will require a lot of resources. We must be honest when we are talking about legislation, people having concerns about a 12-week limit, foetal anomalies, fatal foetal anomalies, incest and rape and so forth and point out that nothing can be done unless we remove Article 40.3.3ofrom the Constitution. All the talk about what might or could happen is only relevant in the event that Article 40.3.3o is removed and the only ones who can remove it are the people. My vote has the same value as that of any person outside the House. It will be the citizens who will decide. Our duty, first and foremost, is to facilitate the

holding of a referendum and establish the commission to observe that it is fair and impartial. Political parties and individuals must ensure, at the very least, that they are honest and up front, whatever view they hold, and that we will all be respectful of each other. As someone who will be 50 years old next Saturday, I do not want to preside over the continuation of the current situation, knowing that next year, the year after and the year after thousands of Irish girls and women will board aeroplanes to seek health care in other countries or climb the stairs to their bedroom with an abortion pill because the State has failed them.

Record id: 4
TD name: Mary Lou McDonald
Date: 17 January 2018

I would like to add my voice to those who have commended the work of the Joint Committee on the Eighth Amendment of the Constitution. I particularly commend my colleagues, Teachtaí Louise O'Reilly and Jonathan O'Brien and Seanadóir Paul Gavan. I also commend the Chairman of the committee, Senator Catherine Noone, who did a very fine job. As an Oireachtas, we made a big ask of our peers and colleagues on the committee. I believe they conducted themselves with great dignity, compassion, intelligence and thoughtfulness throughout the hearings. I commend them on that. I was disappointed that some members of the committee tried unsuccessfully to thwart its work. I do not say that to sound a note of rancour, but because it is important that the disrespectful commentary which featured in the course of the committee's work is not allowed to set the tone for the public debate. As we move to repeal the eighth amendment of the Constitution and agree a new legislative and regulatory framework, I hope and believe it will be an historic and momentous journey for all of us. I hope all of this work can be done in a respectful, well-informed, calm and fair atmosphere.
The harrowing experiences of Joanne Hayes, and all the horrific events surrounding what became known as the Kerry babies scandal of more than 33 years ago, have played themselves out again in heartbreaking detail in the past 24 hours. The scale of the abuse Joanne endured at the hands of the State was unprecedented and horribly and agonisingly public. I welcome yesterday's apology from the Garda Síochána and today's apology from the Taoiseach. Such statements are most welcome. The Taoiseach must now move to make good today's statement on compensation for Joanne Hayes. She is entitled to a full and formal apology from the State for its persecution and vilification of her and her family. She also deserves compensation and redress. Back then, having failed Joanne and having pursued her case in the most corrupt manner, the Garda went viciously for a second bite at the tribunal of inquiry.
I remember hearing Joanne Hayes's name when I was a girl. I remember the Kerry babies being spoken of. I also recall the incredible atmosphere of hostility directed at women and girls in 1983. I can still feel the very toxic atmosphere in which the eighth amendment was conceived, debated and inserted into the Constitution. I do not think there is a woman of my age in

Ireland who cannot still feel how that atmosphere felt. After all, this was the Ireland of the mother and baby homes and the Magdalen laundries. It was an Ireland where women were to be subjugated and kept quiet inside the home to accept their fate. This obsessive control of women did not happen by accident.
It was very much intended by a powerful conservative cohort across society - in government and the churches, across the highest ranks of the public and Civil Service and among the professional elites. Women's subjugation was part of a carving up of power and influence in the public and private spheres, which did not happen by accident.
The then journalist and current European Ombudsman, Emily O'Reilly, wrote the following in 1992:

> The widespread passive acceptance of the patriarchal nature of Irish society also enabled the conservative lobby to hold sway. Nothing threatens the system more than when women are enabled to take control of every aspect of their lives, public and private. And there is nothing more critical to the exercise of that control than the ability to decide how many children to have, if any, and when to have them.

The eighth amendment was in effect a constitutional coup and the reactionary codification of the suppression of women. That is what happened in 1983. In the decades since, women in Ireland have had to live with the abusive outworkings of the eighth amendment. The X case, in particular, brought into sharp focus the worst expression of the conservative coup. A child who was pregnant as a result of rape was dragged through the courts by the State, whose sole and stated intent was to force her to continue with the pregnancy from rape to full term. I still struggle to fully comprehend the callousness of the State in so aggressively and cruelly forcing a child victim of rape to continue with pregnancy. It is hard to fathom that any individual or agency could heap more abuse and trauma on a child who had already been violated. Even after the Supreme Court judgment, the public outcry and the horrors endured by Miss X and her family, successive Governments refused for more than two decades to legislate for the X case. It is important to record that this is a source of shame for successive Oireachtais. It took the tragic death of Savita Halappanavar and the alphabetical array of cases taken by incredibly brave women to shame government into finally enacting in law the Supreme Court decisions in the X case.

There is now broad acceptance across the Oireachtas and in wider society that the eighth amendment must be repealed from the Constitution. I passionately believe that now is the time for leadership and in that regard, I commend the Minister for Health, Deputy Simon Harris, on his words this evening. Leadership must come from the front.

Abortion is a divisive issue. That statement echoes throughout this debate. While that may be the case, the abuse of women and indifference to our health and bodily integrity are not divisive issues but unacceptable positions to take in public life. Some people argue that because abortion is a divisive issue, votes of conscience must, therefore, be allowed. I do not share that view and I say this as someone who is deeply respectful of diverse views and fully understands that some people will struggle with this issue. However, in the final analysis, this debate is a matter of public health, women's health, our right to decide and our right to respect for our conscience as we decide on matters for ourselves. The clinicians and doctors, not least in the hearings of the joint committee, made clear that the eighth amendment casts a long shadow over their practice and relationships with their patients and jeopardises women's health and their lives.

The first issue for the Oireachtas to address is the nature of the question to be put for repeal. A simple repeal of the eighth amendment of the Constitution, as recommended by the joint committee, must be delivered. There can be no equivocation on this by anyone. If legal advice that takes a contrary view is offered, for instance, if it is suggested that instead of repeal simpliciter, an enabling clause should be inserted in the Constitution, the Government must share this advice with all Oireachtas Members because we need transparency and informed debate above all.

I agree with Deputy Billy Kelleher and my party colleague, Deputy Gerry Adams, that the first order of business is the repeal of the eighth amendment. Thereafter, we must debate and acquire an understanding of the legislative framework. The first task and duty of the Oireachtas at this time is to remove the eighth amendment from the Constitution. It is time to right a fundamental wrong that occurred in 1983. As legislators, we cannot accept the terrible impact of the eighth amendment on women's health, their obstetric care and well-being, and their and our fundamental rights. We must state loud and clear that we trust and respect women and that there is no place for the cruelty of the eighth amendment in a modern and diverse Ireland.

There can be no place for unnecessary dogma or doctrinaire positions in the coming months. I accept, however, that we must listen to, acknowledge

and engage with people's concerns. It is, after all, our shared responsibility to protect women's rights and health now and into the future by engaging in a respectful debate and delivering a successful referendum result, which means the repeal of the eighth amendment.

As we are all being confessional and owning up to our ages, I was 14 years old when the eighth amendment to the Constitution was made. As I was reflecting on this debate and listening to Joanne Hayes, I wondered how I would explain to my 14 year old daughter what Ireland was like then and what the hostility experienced by women and girls felt like at the time. I am very happy to say I could not begin to explain to my 14 year old daughter what that was like. That is a great thing. It is now time for the law, politics and every Member of the Oireachtas to catch up with public opinion and the new Ireland, the country in which my 14 year old daughter and all our daughters and granddaughters - and our boys and men - live and give us and them a decent, human rights-based and respectful Constitution that acknowledges women as full and equal people.

Record id: 41
TD name: Michael McGrath
Date: 23 January 2018

I will share my time with Deputy Butler. I welcome the opportunity to make a contribution to this most important of debates. As did other colleagues, I thank all the members of the Citizens' Assembly for their work and contribution and, of course, our Oireachtas colleagues on the joint committee for the painstaking work in which they engaged over a number of months. It certainly was not an easy body of work. We owe them all a debt of gratitude for their commitment and for the sacrifice they made. I know each of them was under considerable pressure to declare positions, and whether we agree or disagree with the final outcome they are to be thanked.
The calls that have been made so far for a respectful debate on the issue have been well made. It is important we all recognise that fundamentally this is a deeply personal issue, and it is a matter for each individual citizen, as part of a referendum, to decide what he or she believes is right. It is important that all of us are non-judgmental in this regard and that we respect the views and opinions of others. I sincerely hope the debate will be conducted along these lines. It will not be conducted like that online. Any Member of the House who declare an opinion or view on this would be well advised not to go on Twitter to see what people are saying about them, irrespective of what position they have taken.
From my point of view, I will support the holding of a referendum. That will involve voting in favour of the referendum Bill. This is because ultimately the Constitution, Bunreacht na hÉireann, is owned by the Irish people and it is a matter for the Irish people to decide what is set out in the Constitution. A significant number of people, we do not know how many, want to see a change to Article 40.3.3°.
I will support the holding of the referendum when that Bill comes before this House.
Our party took the position in 2013 that there would be a freedom of conscience vote and I was one of the people at the time who advocated the approach, which was correct. It will now be replicated as part of this process, wherever it ultimately leads. It is not appropriate that I or any other member of our parliamentary party would seek to impose our views on other members in the party. It is very obvious there are divided opinions within our party, as there are in every strand of Irish society. We have adopted a mature

approach in that respect.
With respect to the recommendations of the joint committee, it is my view that many people - I count myself as one - favour some change but certainly not change along the lines of what the joint committee has recommended. The recommendations go too far and the Government would be making a major mistake if it put the question along the lines of what the committee has recommended. It would constitute a binary or black-and-white choice but many people in our society have a much more nuanced view of the matter. Nobody can say how many such people there are but although they recognise the issues with Article 40.3.3°, they do not support access to unrestricted abortion up to 12 weeks. That is my position. I do not support unrestricted abortion to 12 weeks.
I have read the evidence and legal testimony and presentations given to the committee. I support retaining constitutional protection for the unborn. I appreciate the evidence given by the medics and it must be taken on board in respect of the practical difficulties of differentiating between where a risk to the health of the mother becomes a risk to her life. We have no option but to deal with that and other matters. I favour replacing the existing Article 40.3.3° with constitutional protection for the unborn that permits the Oireachtas to legislate within certain confines. That would not be easy and I am aware of the difficulties that Deputy Shortall referred to in respect of legislating. These are not insurmountable, however, and accepting that they are insurmountable means we would encroach on the most fundamental right of all. The right to life of the unborn should remain in the Constitution but in a practical and workable manner. That is my view.
There is a real risk that if the Government goes down the road it is considering, people in facing the referendum will be confronted with an unknown. If it is intended that Article 40.3.3° is to be repealed, there will be a statement of intent from the Government and I presume there will be heads of a Bill that it would seek to introduce to this House. However, there is no certainty as to the shape of the final legislation that could be adopted after a referendum. We must all accept that. This is a minority Government and its main party has correctly agreed to a freedom of conscience vote. Our party also has a freedom of conscience vote. We have already heard from a number of Deputies in the House who favour repeal and reject the notion that there should be any 12-week limit. That is a reality and they will seek to amend any legislation and extend the 12-week limit to "as late as necessary", which is a term used by a number of them. The people voting on the question of

repeal simpliciter with some form of statement of intent from the Government to introduce a Bill will be very unsure as to what they will ultimately get by way of legislation passed by this House. That is a point that will be consistently made over the course of the campaign.

I do not envy people in making up their minds on this matter. All any of us can do is say what we believe. Nobody can criticise people for stating their own beliefs, based on personal conviction. That is what I will do over the course of a campaign that will inevitably follow in the coming weeks and months. People should engage in an honest, open and respectful debate without judgment. Ultimately, this is a matter for the Irish people to decide. The beauty of our democracy is that my vote, the Minister of State's vote and the Ceann Comhairle's vote in a referendum is of equal value and weight to that of Joe Murphy, Mary McCarthy and every other citizen in our country. That is how it should be.

Appendix C

Getting and using the Rhetorica software

Getting and Using the Rhetorica Software

Rhetorica runs from the Windows (64-bit only) command line. The source code (Visual Studio 2013 solution) resides in a GitHub repository:

https://github.com/priscian/rhetorica

As do the executable files alone:

https://github.com/priscian/rhetorica/raw/master/bin/x64/Debug.zip

Both the VS 2013 solution and the executable rely on external NLP tools:

https://github.com/priscian/nlp

The executable file `Rhetorica.exe` (Windows 64-bit with .NET Framework 4.5.1 installed) requires that the NLP tools repository, which contains files used by the Stanford Parser, OpenNLP, and WordNet, be installed to the root `C:\` directory, so that its path is `C:\NLP\`. If this location is not optimal or possible, then these fields in the file `Rhetorica.exe.config` can be changed from their default values:

```
RootDrive: "C:\"
NlpFolder: "NLP\"
```

If `Rhetorica.exe` is run from the command line without any arguments, it will automatically read in the file `Obama - Inaugural Address (2009).txt`, parse its sentences and find all 14 rhetorical figures. There are two other ways to send a document into Rhetorica for processing:

```
Rhetorica.exe [drive:][path][filename]
Rhetorica.exe [filename]
```

If only the filename is given without an absolute path or one relative to `Rhetorica.exe`, then Rhetorica will look for the file in the directory `C:\NLP\texts\`. For example, the NLP repository's `texts` directory contains the file `obama_2009.txt`, so running the following command will also process President Obama's 2009 Inaugural Address for rhetorical figures:

```
Rhetorica.exe "obama_2009.txt"
```

The command-line interface also allows a second argument with JSON notation for limiting the figures discovered and tweaking the search settings (described in **Table 6**) for any or all the figures; e.g.

```
Rhetorica.exe "Stevens - Farewell to Florida.txt" ^
"{^
  Anadiplosis: { windowSize: 2 },^
  Epizeuxis: { windowSize: 2 },^
  Polysyndeton: { windowSize: 1, extra: 2 },^
  Isocolon: { windowSize: 3, extra: 1 },^
  Oxymoron: { extra: false },^
  All: {}^
}" "stevens"
```

where `"stevens"` is the base filename for the Rhetorica output files. Generally, `Rhetorica.exe` takes three arguments:

`Rhetorica.exe` *`source_file search_params output_pathbase`*

1. *`source_file`*: Path and filename of a text file to process for rhetorical figures.
2. *`search_params`*: (Optional) JSON object with names of the rhetorical figures to find and optional search settings for each figure.
3. *`output_pathbase`*: (Optional) path and partial filename for storing results. *`output_pathbase`* + `.doc.csv` describes each token in the source document, and *`output_pathbase`* + `.csv` describes each figure discovered, in the context of the source document.

Some further examples follow.

Example 1. *Search for all figures in the file* `test.txt`*, then save the results to* `out.doc.csv` *and* `out.csv` *in the current directory.*

```
Rhetorica.exe "test.txt" "" "out"
```

or

```
Rhetorica.exe "test.txt" "{ All: {} }" "out"
```

Example 2. *Search only for isocolon in the file* `test.txt`*.*

```
Rhetorica.exe "test.txt" "{ Isocolon: {} }"
```

Example 3. *Search only for isocolon with a search window of 2 sentences (default 3) and a similarity threshold of 1 (default 0; see § Isocolon for details).*

```
Rhetorica.exe "test.txt" "{ Isocolon: { windowSize: 2, extra: 1 } }"
```

Example 4. *Search for isocolon with tweaked search settings as in the previous example, but then also search for all the remaining figures with their default settings.*

```
Rhetorica.exe "test.txt" ^
"{^
  Isocolon: { windowSize: 2, extra: 1 },^
  All: {}^
}"
```

Example 5. *Search for isocolon with tweaked search settings, and oxymoron with* `greedy:true` *(see § Oxymoron for details), then save the results to* `out.doc.csv` *and* `out.csv` *in the current directory.*

```
Rhetorica.exe "test.txt" ^
"{^
  Isocolon: { windowSize: 2, extra: 1 },^
  Oxymoron: { extra: true }^
}" "out"
```

Example 6. *Similar search to that of the previous example but with tweaked polysyndeton (minimum consecutive sentence-leading conjunctions comprising a polysyndeton = 3 instead of the default 2; see § Polysyndeton), then save the results to* `out.doc.csv` *and* `out.csv` *in the directory* `C:\NLP\texts\`.

```
Rhetorica.exe "test.txt" ^
"{^
  Isocolon: { windowSize: 2, extra: 1 },^
  Oxymoron: { extra: true },^
  Polysyndeton: { extra: 3 }
}" "C:\NLP\texts\out"
```

Appendix D

The Penn Treebank POS tagset

(Marcus, Santorini, and Marcinkiewicz (1993))

Tag	Description	Tag	Description
CC	Coordinating conjunction	TO	to
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential *there*	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present participle
IN	Preposition/subordinating conjunction	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sing. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sing. present
JJS	Adjective, superlative	WDT	*wh*-determiner
LS	List item marker	WP	*wh*-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	*wh*-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol (mathematical or scientific)	”	Right close double quote

TABLE D.1: The Penn Treebank POS Tag Set

Appendix E

Informed consent form for interview participants

TRINITY COLLEGE DUBLIN
INFORMED CONSENT FORM

Lead researcher(s): Emma Clarke

Background of this research
This PhD research aims to derive understanding from research in the field of linguistics to identify a number of rhetorical devices are used in natural speech. These rhetorical devices will subsequently be used to create semi-automated markup, which will have the potential to support NLG research.

Procedures of this study
By agreeing to participate for this research you will be interviewed on the topic of Natural Language Generation. The interview will be semi-structured, meaning that it will take the form of a conversation rather than a rigid question and answer format. The interview will be recorded via audio recording software on a Macbook Pro (Quicktime / Piezo). The recordings and their transcriptions will be anonymised and stored digitally in accordance with the Data Protection Act at Trinity College, Dublin. Each interview is expected to last between a half an hour to 1 hour, depending on participant's availability and the development of the interview. In advance of the interview, you will be provided with some relevant background information which will help to situate this research within the field.

Declaration
- I am 18 years or older and am competent to provide consent.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- I understand that I may stop electronic recordings at any time, and that I may at any time, even subsequent to my participation have such recordings destroyed (except in situations such as above).
- I understand that, subject to the constraints above, no recordings will be replayed in any public forum or made available to any audience other than the current researchers/research team.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- If the research involves viewing materials via a computer monitor, I understand that if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.
- I have received a copy of this agreement.

Participant's name: _____

Participant's signature: _____

Date: _____

Statement of investigator's responsibility
I have explained the nature and purpose of this research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

Investigator's signature: _____

Date: _____

You may contact the researcher, Emma Clarke at any stage of the process: (clarkee8@tcd.ie or 087 7988854)

Appendix F

Information for interview participants

Information sheet for prospective participants

Title

Semi-automated mark-up of rhetorical devices in corpora to support Natural Language Generation research

Background of the research

This PhD research aims to derive understanding from research in the field of linguistics to identify a number of rhetorical devices used in natural speech. These rhetorical devices will subsequently be used to create semi-automated markup, which may have the potential to support NLG research.

Procedures of this study

You have been selected to participate in this study because you indicated in your online research biography that one of your research interests is natural language. By agreeing to participate in this research you will be interviewed on the topic of Natural Language Generation. The interview will be semi-structured, meaning that it will take the form of a conversation rather than a rigid question and answer format. The interview will be recorded via audio recording software. The recordings and their transcriptions will be anonymised and stored digitally in accordance with the Data Protection Act at Trinity College, Dublin. Each interview is expected to last between a half an hour to 1 hour, depending on your availability and the development of the interview. In advance of the interview, you will be provided with some relevant background information which will aim to situate this particular research within the NLG field.

Declarations

- Your participation in this study is voluntary: you have the right to withdraw at any stage in the process without penalty
- The researcher is an ADAPT Centre PhD student. If participating in this research raises a conflict of interest, please indicate this to the researcher.
- In the unlikely event of an inadvertent discovery of illicit activity during the study, these will be reported to appropriate authorities
- No audio recordings will be made available to anyone other than the researcher, nor will any such recordings be replayed in any public forum or presentation of the research.
- Your data will be treated with full confidentiality and if it is published, it will not be identified as yours.

Debriefing arrangements

In order to validate the interview content, you will be asked to review a transcript of the interview before it is analysed for the purposes of the research.

Contact

You may contact the researcher, Emma Clarke, at any stage of the process: (clarkee8@tcd.ie or 087 7988854)

Appendix G

Interview transcripts - round 1

Interview 1
Date: 24 May 2017

Interviewer: Could you describe your own work?
Interviewee: I work mainly in speech signal processing. Analysing speech in terms of voice qualities. If it's expressive, has emotions or voice quality. If it's [UNCLEAR] intense voice. Try to model it, parameters. Also try to generate speech to with that type of speech characteristics, so speech from text. Recordings to extract acoustic features like the pitch that's related with the representation of the signal. These features can be useful then to represent the signal. We can analyse them to differentiate different sounds. Then through these parameters we can then analyse the signal to see for linguistic things in the signal or you can model these features and then use the models to try to generate speech for example from text.
Interviewer: How do you approach NLG production?
Interviewee: So this systems (Text-speech synthesis systems) have different components. It's a bit complex, but you have an analysis part where you extract both analyse the speech signal, also analyse the text. The recordings need to have some descriptions of what is said. So you extract acoustic parameters and then extract linguistic features from the text like the phone(tic) transcriptions. Instead of having characters we can represent the text with the phone units to represent the sounds, the vowels, the consonants. Then you also need to extract other linguistic features. In order to synthesise speech, you need to know about the context – what's the previous syllable. What position is the syllable in the sentence. To see if the syllable is stressed or not. A number of linguistic features. Then see the acoustic context of the sounds within the sentence. Different acoustic characteristics. Then the models – you take both acoustic and linguistic features and try to model / map the linguistic to the acoustics to find that mapping for each representation of the acoustic features. We build those models from a large data set – several hours of speech. Then you use the models to create new text. When you have a new sentence, then you extract the text features from the sentence, the linguistic features. You use that features as input to the system. Then the system using the models that you previously built will generate the test acoustic features. Once you learn the mapping btw the linguistic and acoustics, then to generate those acoustic features. The word 'had' for example, it will generate acoustic features for all short segments along the word had –

phone units: the 'h', the vowel 'ah', the 'd'. What's previous phone before the h, the a, the d because that will affect the form of the ? in the sentence. Then from the acoustic features, using signal processing, you can reconstruct the speech way form from the features. The features are just a compact representation of the speech signal. First, the training stage. It doesn't matter which features you extract first. Once the features have been extracted, they are used to build the models. When creating the models, we first create the linguistic analysis of the text – we don't have the speech now, just the text and we want to generate speech so you just do the linguistic analysis. My approach falls into the statistical ML approach. The deep learning approach. Building these models will be large amounts of text and built using the probabilistic distributions. The first systems were more rule based. The state of the art produces very intelligible natural speech. Commercial models like Google are very high quality because they have a large amount of data that they can build from. They have faster machines. The voices sound like the recordings. If you want to synthesise a voice that's more expressive or if you want to change the voice to sound like someone else, it's more difficult.
Interviewer outlines sample mark up
Interviewer: Do any aspects of your work require / rely on persuasive language?
Interviewer: Not directly. The only voice that ? was speeches from the queen because they're available online. We wanted to see if we can change the aging of the voice. If you can synthesise if the voice is older or younger. It's something interesting. It's a speech style that has some characteristic. There might be things like the intonation prosody that might be relevant or might have these segments. When you are analysing anaphora. There might be some more emphasis / contrast on the repetition. Could be interesting to analyse the pitch.
Interviewer: Could you see an application of this (type of) schema in your work?
Interviewee: Yeah. There'd be different layers. We can have from low level layers with characteristics about the word, syntax etc. This could be a higher layer which would give information about a phrase. This could be inserted as a middle layer. I could see this maybe as kind of another possible level of information that could extend the information about syntax. Also try to use to see if there is syntactic info that is the subject into the modelling linguistic specs and speech inside. And could help to model the pitch/ intonation where to put pauses as it's important to know where to place this so that it

sounds natural. Once the first stage. Do some study to see if these things in speech (anaphora and so on) if they have a correlation with the voice if people change the way they speak when using these devices.
Interviewer: Can you see any potential application (use case) for this schema?
Interviewee: Yeah, if you're focusing on building a voice for political speech, this could be useful. It could be easy to adapt to different contexts. Something interesting could be when generating new content, trying to predict where to put anaphora eg. And which ones to use. The frequency that they occur and some parts of the sentence.
Interviewer: What contextual information would be useful?
Interviewee: I'm not sure. I think that would depend on the application maybe. It depends.
Interviewer: How do you see NLG in your area progressing in the next 5-10 years?
Interviewee: This progresses very quickly with the advances in Deep Learning. There's been great development in terms of generating new content. Generating new text is possible with ML. You could generate a new book by a machine. It appears natural. In this area, machines can do very well in terms of modelling the language. I can see that there are still more developments in terms of applications.
Interviewer: What role do you see interdisciplinary discussion playing in the future of the future of your field?
Interviewee: I don't see much happening with humanities. Maybe for example with computer vision and language are getting more together as they have approaches in common. Also there's applications in our group for creating synthetic voices for avatars - the speech, but also the visuals. Also in terms of psychology – how do people interact with avatars. Expressiveness... Higher quality means that people interact more. Try to match expressiveness of the character with expressiveness of the voice. Interplay. I think with humanities, there can be an advantage also, especially when you can work to render more expressive content – in arts / movies. It's very challenging.
Interviewer: Anything else?
Interviewee: I think there's a value for the field for looking at these devices. I think it's going to be interesting as a research area to see how it is, how the linguistic mark up could reflect the voice. You could use it to put more variability in the voice. How to better predict intonation. This could be even more relevant if the focus is on particular speech.

Interview 2
Date: 25 May 2017

Interviewer: Could you describe your own work?
Interviewee: I started off as a mechanical engineer and then I spent a lot of time in 2nd language provision, particularly in a campus company here where we designed refugee language support so I'm also from a linguistics background as well as a more technical background. During that time, I also did a masters in [COURSE], the MPhil here. Then, late in life, I decided to come back and do a PhD. I also did a postgrad diploma in [COURSE]. My interest has always been dialogue. At that stage, there was nobody really doing spoken dialogue in [PLACE] really. I was always interested in inter-speaker effects in dialogue. So what I looked at there was syntactic and lexical priming. My interest was what is really priming but the buzzword tends to be either accommodation or communicative adaptivity where one person's lexis, syntax, choice of rhetorical devices probably would aswel would influence the others and we can see, we've used authorship identification to see over the course of a dialogue to see if the two people's language is getting closer. Now it's at the level of letter counts, so it's way way below the level of doing a tagger or a parser. Which works, if someone's language is that close to another person's, they'll have the same letters, the same frequency of letters. So that was between Russian learners of English and native speakers of English. So that was Masters. And then I came in here because my interest then turned to, and this is where we differ, I suppose it's kind of pragmatics really, it's where we would talk about speech, both the timing, the prosody and how that gives shape to a conversation. So what I concentrate now is casual, totally unscripted conversation and if you like, the superstructure. If you ask most people what happens in a conversation, they think that people are always talking over and back, but they're not. If it goes more than 5 minutes, you get these long chunks of about half a minute and then more interactive chat where one person basically tells a story and then they [UNCLEAR]. And I'm just looking at the temporal information, the distributions of laughter and there's fluency aswel. So there's very little text in the sense of written text and spoken in my current work. I don't go into what is said, more how it is said.
Interviewer: How do you approach NLG production?
Interviewee: There isn't strictly generation in my work at the moment. Most conversational agents at the moment, unless they're doing a very standard

task. The classic is... The sad fact is that the field grew. The challenges and the grand challenges in the field for years was something that would tell you railway timetables or something you could book a flight from. And they're so standardised that you can make it seem like you could use a stochastic process or a scripting process to make it seem like you're doing generation, but the number of slots is so narrow, you're not. Whereas with casual conversation, you could argue that it's not what you say, it's how you say it and there seems to be a motive rather to entertain and fill the time than to get anything done. So at the moment we're scripting most of our, we build systems that kinda chat to you and they're scripted at the moment. And most NLG in those systems is either completely stochastic – it just pulls items from a database of phrases that have worked before or ngrams or chunks that have worked before or they're scripted.
Interviewer: Could you outline some limitations of current NLG approaches?
Interviewer outlines sample mark up
Interviewer: Do any aspects of your work require / rely on persuasive language?
Interviewee: Anything that will allow a formalisation of the kind of stuff that happens in spoken language. Because this text was written to be spoken, it is a speech genre. Everyone wants to generate spoken content. And it's small enough to be of value. You're not saying "I'll give you something that will make speeches" – you're saying "I'll give you patterns that can add something". The only thing I'd worry about. There's something..... It's actually very interesting when you start looking at it. Have you looked at any of the priming work? If you go into psycholinguistics rather than rhetoric.
Interviewer: Anything else?
Interviewee: If you have anaphora like these, cos a lot of these anaphora are syntactic, there is very little lexical. They're glue rather than stones if you know what I mean. Because of that, they would be quite easy to pick out so you can search pretty easily for them. What is the mark up adding? How are you sure that it's actually anaphoric? The difference between "I'll go to the cinema and then I'll go to the shop and then ...." and "we will do this and we will do that" needs to be thought out. What exactly is the meaning if the rhetorical device rather than the meaning of the words? If you get that, it's very useful. You need to formulate something that the machine cannot do at the moment. What do these devices bring. These speeches are written to be spoken. They are also in a sense written to be read but they're written to be spoken and the devices within them, although it's not coded into

the speech, there are prosodic devices in there. The different between "We will fight them on the beaches and we will . . . " and "we'll go to the cinema and we'll . . . .." – that's a prosodic difference, but if you read this speech and you're rehearsing it in your head, the rhetorical device will come to you.

Interview 3
Date: 20 July 2017

Interviewer: Could you describe your own work?
Interviewee: I guess you could summarise it as different aspects of lexical semantics or processing lexical units in text. For example the paper I am working on at the moment is for multi-word expressions. It is about identifying multi-word expressions, particularly verbal multi-word expressions. Multi-word expressions that are formed and that involve a verb. So it can be phrasal verbs for example – look up, look down, put up with etc. Or it could be idiomatic expressions like kick the bucket – not compositional etc. So it's about detecting them. Or it could also be there's a newer theory of recurring verbs or verb phrases that are not phrasal verbs and are not an idiom – the example they give is 'to have a conversation' they are called unclear live verb? constructions. Have is considered a weak verb in the sense that it can mean a lot. The meaning of the phrase 'to have a conversation' comes from the noun, rather than from the verb. It's a phrasal verb – a phrase and a verb – a multi-word expression but the main meaning comes mostly from the noun, the object in this case and then the verb tells you the type of action that we're going to have. That's what we're working on. What I do in this type of examples is I try to apply Machine Learning – syntactic features and semantic features and the paper we are working on uses a Conditional Random Field CRF model which is used a lot in NER and basically goes through a list of tokens, a sentence tokenised and just goes into which tokens form the multi-word expressions and which ones don't. And we just extract that. We also use word vectors as in sematic vectors to try to help the CRF algorithm. We wanted to include it as a feature too but we didn't have time so what we decided to do was let the CRF algorithm do its work and normally these algorithms when they are used, people get the most probable sequence of a sentence and it will come back and tell you, it will tag each word in a sentence and tell you that this word does not belong to a multi-word expression or this one does and this one does two together. This is called a sequence link – a sequence that is labelled. We try to, the CRF algorithm gives you the one sequence of labels that is more likely, but it can actually compute several candidates – so what we did is we asked the algorithm to give us the ten best (most likely) candidates according to its model. Then we modelled those candidates as semantic vectors from EuroParl then we computed cosine similarity which are used a lot in lexical semantics to try to detect a

bi-proxy. Whether a multi-word expression is compositional or not compositional. We expect that eg. in the phrase kick the bucket, that the overall meaning 'to die' is very different from the individual meanings of each of the words. So we use that as another feature til we ranked the ten best outcomes of the CRF. So we observed some improvements. So that's what this paper is about. My background was in terminology in the IT domain specifically. Where I am interested in the specialised language and how it can be used in machine translation and information retrieval among other things. I am also interested in terminology and the lexicographic process itself. You know if you are detecting multi-word expressions – why are you detecting them? Presumably you are compiling a dictionary of them. Or you are trying to improve a machine translation system that will. Or if you feed them in some way to a system, they might improve the quality. So it's everything to do with words and terms and how they fit into other processes.
Interviewer outlines sample mark up
Interviewer: Do you do mark-up in your work?
Interviewee: In NLP in general, we do have a lot of mark-up problems. In NLP usually when you download a corpus that is annotated, usually people in NLP when they see very complex annotation, they go 'urgh' and they either convert it into something simpler with a perl or python script so they can process it quicker. My multi-word expressions thing – they are broken down one word per line. Every token appears as one column. Usually you get the part of speech of the word and then you may add many other features depending on what processing you use and then you would have the label that would tell you "this is not a multi-word expression", "this is not a multi-word expression", "this is the beginning of a multi-word expression", "this is an intermediate word of a multi-word expression" and "this is the final word" for example. Basically, when I use an algorithm, the algorithm uses these things as input and this is the output – this is what it is going to predict basically. During the learning process, I tell it : use this column as the input and this is what you have to learn. And then when I present it with another sentence, it's able to predict the same multi-word expression. The one I am using is PARSEME. It's basically the people from the COST action who came together to develop that. This in turn is based on the CoNNL format – CSV columns. CONNUL from the universal dependency project. It's a bit like... There have been quite a few efforts to build parsers for many languages. The problem is that you start working on Irish for example or Spanish or French. The teams start working together and they come up with POS tags that are

incompatible. When you are talking about languages that are similar, for example French and Spanish, you could easily share the .... The syntax is largely the same you know. Many languages can share so this. What they are trying to do is homogenise all the languages and trying to convert all these different tag sets into a common one. And this columnar format, this is what they use. They haven't chosen an XML format. For some reason, a lot of people in the NLP world, they like the simplest possible. There is a lot of work on anaphora resolution. I'm not sure if people in that area have come together and tried to standardise a markup format or an annotation format. It usually happens that.... We participate in a shared task. The good thing is that people get busy and come up with a human annotated corpus and you just have to write a program to use it. So it saves you time. When people design these things, they have to make design decisions. Sometimes you realise they didn't do the best ... usually there are several iterations, they improve, the annotation improves, there is disagreement between people. Especially with semantics and things like that. Not so much syntax. For example what is a multi-word expression? May have a disagreement. Many times with these shared tasks because it's a practical thing, they have to come up with a format to share the files and everybody has to use the same format. That is a problem that how are you actually going to go about that? In the very early .... You know the problem with word sense disambiguation. SemEval looked at many types of evaluation. Look at the type of formats that they are using. I think that these people at SemEval they are doing stuff in anaphora so it's certainly worth a look. The original shared task was called SensEval. These people changed it to SemEval. Earlier it was specifically about word senses. When I was doing my PhD I was working with their original corpora. They started in 97 or 98. And that's when XML was only taking off and they were like "let's try this XML thing" but actually if you are very technical and you look at it. It's not XML compliant. It looks like what you are doing here but if you put that in a parser, it would be rejected. So they ... There's also very serious efforts at really having a common linguistics approach to this. At the end of the day, there is loads of standards, loads of initiatives and not that much consensus but if you want to know more about this, I would look at the SemEval shared task on Anaphora. Nowadays, SemEval competition is 14-20 shared tasks usually on multi-lingual translation etc. There's an explosion now of frameworks. NLTK is one, but there are others coming up. There's one called Spacy actually – it's also python. They try to be a bit cleaner than NLTK (which is a bit of a mess).

Interviewer: Where do you get your corpora?
Interviewee: One way is from the shared tasks. There's a few repositories of corpora. There's a lot of initiative to provide stuff for free. There's also ELRA. A lot of these formats come from different communities.
Interviewer: Do you think that these formats can capture nuance / subtleties? Context?
Interviewee: Yes and no. What people do is they say I want to add another feature, add another column and another column. That's what people do and leave the Machine Learning component to the whatever. Sometimes people can do that. Sometimes people are very crude.

Interview 4
Date: 02 August 2017

Interviewer: Could you describe your own work?
Interviewee: My main research topic is about Machine Translation (MT). I do some research and also I get involved with some industry collaborative projects. So for the research I mainly work on the algorithm or the methods of MT and Machine Learning (ML) methods. And currently we use the deep learning methods, specifically the deep neural networks for MT. In terms of collaboration with industry projects we actually integrate the MT system into their product pipeline for translators. And there. For example, for the project with [COMPANY NAME], the [COMPANY NAME] Lab in [CITY], it's a dialogue based MT system. A dialogue MT. But we don't work on speech recognition or speech synthesis. We normally work on text translation from dialogue MT systems. Do for the [PROJECT NAME], we used some companies ASR and speech synthesis for Chinese and English. But MT engine is developed by us. In the [PROJECT NAME] project, we used two [PROJECT NAME]s. One of which is a reception in a hotel in [COUNTRY] and the other is a customer tourist in Ireland and she or he wants to travel in [COUNTRY] but both can only speak one language. The receptionist in [COUNTRY] can only speak Chinese and the tourist can only speak English. We use the MT system and speech recognition and speech synthesis together to help them to finish a conversation such as a hotel booking.
Interviewer: Is it all done with deep learning or is there a mix of Deep Learning (DL) and rule-based / template?
Interviewer: Currently for the [PROJECT NAME] demo we don't use too much DL methods because for the spoken language, the conversation. First of all, we don't have too much training data. For DL methods, it relies on large scale datasets so it can achieve a good performance. But for small scale datasets, currently the neural network methods cannot perform very well. So for this demo we still use statistical based methods for MT. And also in the semantic model which can help the system understand the context of the conversation. And we use combined methods. That means we still use some manually defined rules to understand the question or the sentence in the conversation and also we use statistical based methods to understand the semantics of the conversation so it's a combined model for the understanding of the conversation. For the MT we use purely statistical based methods which is the classical ML methods rather than the DL method.

Interviewer: How do you approach NLG production?
Interviewee: I think the final goal of MT is to generate natural language. So but currently we still have some problems in the generated language. Non-natural words or errors in the translation. For example word orders. Humans can actually see the language in a good word order, but for our translation the main problem for us is the word order especially when the sentence generated is very long or there is a long distance dependency in the words in a very long sentence. Currently the MT system cannot perform this well. Secondly, and we also have some grammatical problem – for example the tense – the past tense or the present. The MT cannot perform very well. But it also depends on the use case. For some very specific domains, if we have enough data then we can train pretty well a MT system. And the generated sentence is understandable and even if there are some errors in the word order. Even if there are some errors, a human can guess the meaning in the sentence, but for machine to machine it is still a problem.
Interviewer outlines sample mark up
Interviewer: Could you see an application of this (type of) schema in your work?
Interviewee: I think for this. These are useful for MT. There are a lot of errors in the translations. For example, for anaphora and in the source sign and also in the target side. And we also need to translate as anaphora. The anaphora is definitely persuasive. It is very important for a MT system to keep this in the target language. But, in the real MT systems we cannot guarantee that this is kept in the translation because of the word order or something like that. And also, here "it was the" but we cannot guarantee in the target side this will be translated consistently. It might be "it was" in the first one but in the second one it might change to "it has been" for example. But if we have this markup, then we can use this at least as a feature in the MT system. That the system can recognise, ok this is for example anaphora and then we can keep the format in the source sentence and then keep this in the target sentence. I think this is useful because actually we have a lot of similar work in MT with this and we need to recognise for example the relationship of the sentence we are translating and we need to recognise the relationship of this sentence with the previous sentence and also the following sentence and it might be like ... The previous sentence might be the reason and the following might be the result. This is called Discourse MT. So we need to keep the context and the relations between the sentences so we can translate maybe the paragraph or this document consistently. Rather than we just translate

each sentence independently. So I think this is definitely useful. This can provide context information for the translation.
Interviewer: what would a useful format look like for you?
Interviewee: I think this format is ok for us. This is an ok format for us, we can process this one. We can actually, we use this information as a feature and whatever format it is, finally when we use it, we need to extract the relationship and represent this relationship and feature. The kind of format doesn't matter for us because we can't directly use this format. We need to convert the format to a format we can use in a MT system. I think this ok because it clearly marks up this anaphora and this is symploce and we can convert this to a format we can use in a MT system. And also in MT or in the translation industry we also have an XML format dataset which is called TMX – exchangeable translation memory. So for example when you finish the translation of a document, you can store the translated documents in a database, but the data is stored in XML format and later the translator can use this translated document. If for example, when the translator translates another document, which has some similar sentences in the new document and the already translated document sometimes they can retrieve the already translated sentences in the TMX document and directly use them and just post-edit some different words and reduces working time and improves productivity. In TMX actually we have some more complete markup than this – for example this is bold words and there is a mark up need to see that these two words are bold. For this markup and when the TMX file needs to be displayed using a browser, it should show these two words as bold. TMX is often used in localisation industry. For example in computer games. Originally computer games were written in English and often some buttons have English words. But now, this product needs to be sold in [COUNTRY] so this word needs to be translated in [LANGUAGE] but for some words in a button, they might be in bold or in italics. This font properties needs to be kept in the translation so this is kept by the markup.
Interviewer: What role do you see interdisciplinary discussion playing in the future of the future of your field?
Interviewee: For some in MT they have interdisciplinary discussions but for me not really. They are getting more work on MT and ethics.
Interviewer: Anything else?
Interviewee: I think this is very useful, especially in the spoken language translation, speech. Because we need to keep this in the translation so I think it is very interesting and useful.

Interviewer: brings up UN translation – is that MT or live translation?
Interviewee: Currently machines can't do that type of translation very well. I think it's human translation! I think this is very interesting and useful because in MT we also have a similar work about how to keep sentiment in the translation. So suppose in the source sentence there is a certain sentiment and we need to keep the same sentiment in the target. And also we need to detect what kind of sentiment is in the sentence and we need to.... So the sentiment will help us to select the correct words or the phrase in the target sentence. Then we can keep them consistent. For gender – eg in French, different words for gender. In a target translation we also need to select the correct words. Gender can also be used as translation feature. So MT is not only about mathematical models, it also considers a lot of linguistic things and also some social things like human gender, things like that.

Appendix H

Interview transcripts - round 2

Interview 1
20 November 2018

Interviewer: Is this something that you can work with?
Interviewee: It's difficult. Like last time I think we said that. I definitely think that if we have rhetorical type of language, it's a different style. There has been quite some work in MT on different levels of politeness which is not the same, but could be in some way comparable. Like you add a tag in the beginning in some way – like OK, we want it to be very polite or we want it not so much to be polite. I think you could have a similar thing saying like this is rhetorical style but then particularly for those phenomena, that I don't know. I guess it could help, but I don't know exactly how to… how you would integrate the information.
Interviewer shows standoff XML markup to interviewee.
Interviewer: So, something like this?
Interviewee: Something like that? I wouldn't know how to integrate something like that. Particularly, because usually we work… I guess with something like this, we can get information at the word level. Because what we usually do, if we … Or maybe we could like…
Interviewer: So, I guess one of the things I was thinking of was, rather than doing manual markup, because it takes a long time, would be to mark these or find a way of marking these, and displaying them so that maybe they'd show up in a post-editing process.
Interviewee: aha, that would be interesting I think. I mean because you would want the "it was the" to be translated the same all the time. In MT, although it has shown that it is quite consistent because it relies on statistics, still this is important to keep.
Interviewer: so if there was a way of showing these consistencies or that these are being used in a consistent way, that they may be being used in a specific way. Because we don't know either. These examples are famous examples, but in everyday speech, we don't know whether these devices are being used intentionally/deliberately or not or whether they are being used subconsciously. We can't say for certain.
Interviewee: Did you look at automatic translations of these famous speeches? It would be nice to look at whether the translations work for these.
Interviewer (summarised): I did, but I didn't bring it today. I've done it looking at a Spanish translation and it didn't maintain consistency. For example, in Spanish I = yo, but it is not necessary to use Yo always in the way that we

would use I.
Interviewee: I think it's really interesting. But like for example, there is no way currently, or maybe there is, but I haven't seen any way that you could say: now you want yo to be there, but now you don't want it to be. Although maybe there should be. It doesn't always get translated, but only in certain cases, when you want to put emphasis on it.
Interviewer: It's that particular kind of nuance that I am looking at where if it was a Spanish politician delivering a speech, maybe they would use the Yo for that emphasis, whereas you wouldn't in everyday speech.
Interviewee: Yeah, and then you would want it for all of them.
Interviewer: And maybe it's just a case of marking those in a way that it draws a posteditors attention to it?
Interviewee: Yeah. Or in a pre-processing way. You could mark it and hope that the system learns by itself.
Interviewer: OK, can it do that? For a deep learning translation?
Interviewee: The thing is that it is so hard. This is something that is quite advanced and MT is not yet at the point where this is something that we can work on now. I think it's definitely something that we should work on at some point, but now there are still other issues which are not solved so it's gonna be hard to. And on top of that, if you integrate features marking these things, then it's also very hard to evaluate. It's hard to know what it has learned, what it has not learned because it is this black box so I have similar problems. I am trying to integrate also meta information like "Emma is speaking and she is a woman, so we need the endings to be female.". But there, it learns it sometimes, and I don't know when it is going to be correct and when it's not. It's definitely better with the features.
Interviewer: So, if this was treated like a feature, is there a...?
Interviewee: The work on politeness adds a tag like Polite or Not Polite in front of every sentence. So you could do something like 'rhetorical' or 'non-rhetorical' but that would be on a whole sentence level.
Interviewer: And is that part of the pre-processing?
Interviewee: That is pre-processing. You would have to take this sentence, run your tool. Is it a rhetoric sentence? If yes, add the tag, if not, not. And then train it like that. A second option is like:
.....the |R people|R ..... , ...the |R people|R .....
On every word, attach some tag saying this is a rhetorical structure.
Interviewer: So, you might say, say "this issue", you would add a tag to that?
Interviewee: Then you would add a tag to 'this' and 'issue' saying that they

belong together. But again, I worked with these kind of features and I got always small improvements, but I don't know where they come from. I added a lot of extra information. I had no clue what actually improved. We don't do the evaluation manually. I always do a little bit of it and try and look at it, but it's hard to say.
Interviewer: OK, is that using Bleu Scores?
Interviewee: yeah.
Interviewer: so maybe the way to go with this for the moment, is to have it showing up for a post editor? Potentially. Maybe that's where the value is now? And detecting them is future work?
Interviewee: Ya, because here of course, you would need translations. You would need say the English and the French. It doesn't need to be marked up but you still need a translation and if that translation is not good (I mean it happens!), then definitely, it won't learn these kinds of things, but if you are working on post editing, then you only need the input text. Because then you would just use the input text to predict how "rhetoric" some sentences are and then mark it up somehow to say "rhetoric speech here". I mean, I think it could. The problem is always: if you want to train a NMT system you need 2million parallel sentences.
Interviewer: outlines that one of the key things in the literature on this subject is that a LOT of annotated data is needed before RDs can make use of of ML resources.
Interviewee: So this tool does of all that tagging automatically?
Interviewer: Ya, well it pulls out all of the information but sometimes it has. I don't use all of them. Take the example of (ep)anaphora. If it's just the repetition of 'to' and 'to'. I just use epanaphora which are two or more.
Interviewee: Did you develop this tool?
Interviewer: No.
Interviewee: It doesn't seem very smart. I mean the least they could have done is taken into account the frequency of the words. Then you would not have this..... but then again 'we', 'we', 'we' – that would be a frequent word. But the repeats are still important somehow. But 'the' 'the' 'the' or 'to' 'to' 'to'.... Oh but in some cases it could be.
Interviewer: Exactly.
Interviewee: Argh, it's very difficult.
Interviewer: And when I don't include those 'to' 'to' 'to' or 'I' 'I' 'I', maybe I'm losing something, but at the same time, I think in order to make the point here, it's enough to go with two or more.

Interviewee: And if you include that many constructions, then every sentence is going to be rhetoric.
Interviewer: Exactly. Whereas here, this is only 4 devices. The tool actually pulls out 15 but I figure if I can get something that works for these, but they need a different device, they can use the same process. That's why I've just limited it to 4 here.
Interviewee: So, what do you need to continue?
Interviewer: Well, I suppose, the purpose of the markup and talking to people .. this is like a case study. Having these conversations. I think this is valuable and talking to people with expertise is valuable for me.
Interviewee: I think you're lucky talking to me, because if you talk to one of the more hardcore computer scientists that have zero linguistic knowledge or interest, then you would get the answer: "the machine will learn it" but they think that about everything. Many people are like that. Even the stuff that I do, people say "why do you bother? The machine will learn it by itself"
Interviewer: Will it though, do you think?
Interviewee: I always thought no, but the longer I'm in the field, the more I think it's getting really better at stuff. Do you know [NAME]? S/he always said, "the human translators are OK, we're not threatening their jobs", but the last time I talked to them, s/he said "you know, we are"
Interviewer: Really? Wow, OK.
Interviewee: ya, I think the issues in our field will be more like these issues – it's becoming harder and harder to find little things, nuances to fix. I think they will get there to thinking that these need to be looked at. It's coming I think.
Interviewer: I guess this is language skill at its peak in a way, being able to construct things in a way that is memorable and persuasive. Not every human can do this either.
Interviewee: that's true, but for some things, machines are kinda. For example, if you talk about consistency and stuff, there is quite a bit of work on how to keep a translation of a particular word consistent over a document because if you have a document talking about "some word", you don't want it if it's an ambiguous word, you don't want it to be translated in one way and another. And they actually showed that MT is more consistent than human. It might be the wrong translation, but it will be consistent in its mistake. In some way, maybe for a machine this is easier. But then, these things are interesting polyptoton.
Interviewer: I suppose the thing about these three (epanaphora, epistrophe

and polysyndeton) is that they are direct repetition, but these three are different.
Interviewee: ya, and this is really playing with language here polyptoton.
Interviewer: it gets even more like – here I would not have picked out incisive as a cognate of decisive.
Interviewee: it does look like this is done on purpose.
Interviewer: ya, I think it is.... Yeah.
Interviewee: and then there's a question of whether you can keep this kinda thing in the other language.
Interviewer: presents study related to this – Smith (2006)
Interviewee: I don't think a machine will learn that!
Interviewer: I don't either, but I mean I don't know!
Interviewee: I would probably not be able to work with this sized dataset. That's the issue with a lot of stuff I want to do. There just isn't the data. Most baseline systems work on a sentence level.
Interviewer: So, it won't even learn that – this sentence (1) is connected rhetorically with that one (2)?
Interviewee: What happens internally is that all sentences get shuffled so it wouldn't be able to learn unless you give it that particular sentence as a context.
Interviewer: But if you gave both?
Interviewee: yes, if you say: "this Is a rhetoric block", then you could somehow. You can do that.
Interviewer: OK, so what I have here (shows XML standoff markup)
Interviewee: yes, you could link them somehow, but things that repeat like "I was" and "I was", the machine is quite good at doing those systematically. The other things like the publicity stuff (Smith, 2006). That is more interesting. Stuff like alliterations etc. These things would be very interesting because there you definitely want to keep that somehow.
Interviewer: These polyptoton's are similar to that too.
Interviewee: yeah, I mean that's what I think.
Interviewer: OK.

Interview 2
20 November 2018

Interviewer: What's your own area that you work on?
Interviewee: I work on [LANGUAGE] language machine translation. Usually with English as the source language and [LANGUAGE] as the target. Part of my funding comes from the [ORGANISATION NAME] and they use machine translation… so it's improving the machine translation that they use in the department. But my background is 50-50 linguist and computer scientist and that … I'm not a computer scientist.
Interviewer: I'm not at all. I'm similar, well not similar but my background is … I did French and German at college and then a Masters in Applied Linguistics and then another one in Digital Humanities and I've ended up in the computer science department in Trinity. Where I like working is on the linguistic side of things. I did request some speeches, one in particular from An Taoiseach's office but they haven't got the translation to me yet and that was months ago…..
Interviewee: Were you looking for [LANGUAGE] translations?
Interviewer: Yes, because I was kind of thinking of looking to see how say the Taoiseach's speech, and it was a speech that was delivered in the UN, so it would have been crafted and these devices would have been present (because they are always present even if they are used subconsciously so I was going to look at how these might have looked in a translation.
Interviewee: That would be so interesting to see if these kinds of phenomena happen in Irish as well as in English and would the translator use the same repetition…
Interviewer: Exactly, I guess that kind of one of my things that I'm looking at – how might this be of value and I think that one of the ways that it could be of value to somebody working in machine translation would be to use these, to mark these in a text, possibly for a post-editor to draw their attention to it so rather than actually trying to tell the machine how to translate it because from talking to [NAME] and people, it's not as easy as just saying "do it". To actually mark it like this shouldn't be so difficult.
Interviewee: I can definitely see its importance from a post-editing point of view. Clearly these are used for emphasis and for certain reasons that isn't present in the translation. And then I guess it doesn't have the same effect. So if the translator was flagged – look whatever word you choose here, you should repeat that word.

Interviewer: And in terms of trying to tell say a system that you work with, trying to tell it?
Interviewee: Yeah, that's definitely more difficult to do. Especially because most people have moved on to these neural models and we call it a black box – you can't really delve in and change things as easily. But there are post-editing, not in a human sense but in a machine post-editing or automatic post-editing that it could be a venue I guess for you to consider. So if you had it flagged in the input and then if you looked at the output and you didn't see any repetition there, maybe that could be an indication that your translation isn't great. We tend to use stuff like that for [LANGUAGE] because we don't have this huge amount of data that other languages have so you have to be a bit hacky and a bit crafty about getting translations up to scratch. So we have some type of post-editing at the moment so I can kind of envision that.... And it tags the specific words rather than the sentences? Or does it just tag the sentences?
Interviewer: It tags the words. It would tag. I guess this detects them and gives an output. There isn't a tagger as such. This is me the manual markup. I actually think that this wouldn't be so difficult to automate. I'm not going to do it as part of my work because ...
Interviewee: That's just huge
Interviewer: yeah, and I couldn't do it anyway even if I wanted to! But I guess, this is just a part of, this is like a case study. My PhD is more about having interdisciplinary conversations and you know. Because a lot of the time in digital humanities work, people might create a tool or say this is very useful or this has a value, which it does in a humanist sense, but it's only when you actually talk to the people who use this stuff and hear ... So it's about having those conversations rather than actually figuring this out.
Interviewee: This is a huge challenge.
Interviewer: Yeah, and even a lot of the literature around these devices is saying – and there's actually not that many people who work on these – it's saying that one of the problems with this is that for centuries we've had the same examples – you know like this is Dickens, this is Lincoln's Gettysburg address. We don't have a huge bank of examples even. So a lot of work is needed to annotate these within corpora in order for them to be useful because ...
Interviewee: That's really interesting. You might be able to find if there's any speeches that the Minister for [PLACE] has given – that should definitely be in [LANGUAGE] aswel.

Interviewer: I'll have a look. They'd be on their website would they?
Interviewee: They should be...
Interviewer: Actually, you've just given me an idea – during Seachtain na Gaeilge, they deliver their speeches in Irish in Dáil Éireann. Let me have a look.
Interviewee: And it's not so common, but there must be some [LANGUAGE] spoken in say like the European Parliament. They have to have interpreters there just in case.
Interviewer: Yeah, cos I was looking. I just assumed, maybe naively I suppose because where are they going to get the resources for all of this. That everything that everything that is said in European Parliament is automatically translated into all the languages of the Parliament but that's not the case.
Interviewee: right now, [LANGUAGE] is an official EU language, but when it was made an official language in I think it was 2007, they said that they didn't have the translators to be able to translate absolutely everything so they put a derogation on until 2022. That means that very little will be translated only really important things. So once 2022 happens, they have to translate absolutely everything. But right now, we don't (probably still won't by then) have the translators because people tend... there's even very few courses for it, like Masters for it. And people who do [LANGUAGE] tend to really like [COUNTRY] and don't want to go to [PLACE].
Interviewer: so will they move towards machine translation then? Automating it?
Interviewee: Yes, so that's part of the reason for my work aswel. To try and get the MT up to a point that we don't need loads and loads of translators.
Interviewer: and that's quite soon!
Interviewee: it is! I'm due to finish in [YEAR]!
Interviewer: that's brilliant
Interviewee: it is really interesting work. And I often, it's a real Irishism I find repetition aswel. We looked at it a little bit in Hiberno English where we would say "big fat rat" but they both kind of mean the same thing. You're not going to have a small fat rat or a big skinny rat, but I would imagine that it would come up in Irish aswel but I would love to see if it does.
Interviewer: I never even thought about that – that we might be even more repetitive because of Hiberno-English. But I guess that's beyond what I'm looking at right now. But even, and this is purely anecdotal, but one of the people I am looking at in my speeches now is [TD name] and they use way

more repetition and I don't know if it's because they tend to waffle a bit more, but they use way more devices for repetition.
Interviewee: Is it part of the country they are from?
Interviewer: potentially
Interviewee: maybe more Irish influence?
Interviewer: potentially, but obviously I won't be mentioning that in my work because how can you qualify it? You can't say "they're from wherever..."
Interviewee: it's really interesting. It's really cool that this type of research is being done.

Interview 3
29 November 2018

Interviewer: Would something like this be useful in the work that you do?
Interviewee: I know that anaphora resolution is one of the things that machine translation still struggles a lot to do. Because of the like for example, the example you have it was the best of times, it was the worst of times is a neutral pronoun. It is like if you had a proper name here like Maria, no that is not good because in English you have. If you have for example, the beer was the best and then it was da da da. In English you keep the neutral but if you translate into other languages, you need to know if it is feminine or masculine right? So anaphora resolution is difficult for MT. But I don't know. For the implementation of MT, I don't know how… For the post-editing side, for the evaluation side, this could be useful if you could highlight this. If you are looking to specific things you know so for example if someone tries to implement something for anaphora resolution in MT and then they give me the output of the MT and I know that specifically I have to look at anaphora, having it highlighted in the text so that I can focus on that, Yes. For any of this would be.
Interviewer: For any of the devices. Yeah, so say for example if you are translating this. It's a statement that was delivered in Dáil Éireann and we know that its purpose was to be persuasive or to make a point clearly and if we are saying this language is important for persuasion. And OK, we don't know for definite that these things were chosen deliberately, but the likelihood is that they were, so if your attention was called to these devices, would that be?
Interviewee: Exactly, So neuro machine translation right now, it's a little bit more creative that statistical machine translation. So there is a chance that neural MT will try to use different terminology for this and this 2 different 'I was' components of Billy Keller's anaphora number 125 so it could be a way of enforcing an MT to keep with the same terminology. It wouldn't be very useful for SMT I think because SMT kept the consistency. It would go for the most spoken one. But in a post-editing set up, to have that highlighted and say… Because we have two types of post-editing, we have the light post-editing where we ask post-editors to keep as much as they can from the source so they wouldn't change these things but there are some translators that think it's too much repetitive so they would try to. So yeah, lighting it and enforcing them to like you have to keep this because this the word, the

intention of the speech is coming from. Yes. Especially this also this and, and, and, and, here, they would probably try to get rid of it because too much repetition, unless it's something like a poem that you know you have to but that's more creative translation. I think it's a little bit outside from what Natural Language Processing. So yeah, the first thing when I look, when I see post-editing is that if I have to either enforce translator to translate like that or ask them to specifically look at that to see. So for examine if you know someone builds a system for anaphora resolution in NMT and they said I want you to check if my anaphora is working so then if I have that highlighted, it is easier for me to judge and then I can assign a score of one and zero of correct and incorrect and do an annotation for that so then I can feed back that to the engineer and then he will try to do that. But for that I can only think of anaphora, but for post-editing if you want to ensure translators will keep what you need then everything (here they mean eg all four devices presented at start) but for machine translation implementation and feedback I can think only of anaphora because I think it is the most common one. Because this one, machine translation would translate anyway. I don't know to be honest if they would try (reads from example of polyptoton – strong – strength; skill – skilful etc) . No, I think it would do depending on the language pair, it would.
Interviewer: It would translate them exactly, would it?
Interviewee: Yeah
Interviewer: I might even see how Google deals with a few of these examples maybe.
Interviewee: You could try that. I think they would try to keep consistent. But anaphora, if you look for anaphora resolution for machine translation you'll see that it is a big problem. I don't know if you know [NAME], well s/he works with gender in machine translation and it's basically anaphora resolution because you don't know what each is referring to – female or male. So there is a big case for anaphora in MT.
Interviewer: ok, and then in terms of post-editing tools that you use. Is this something that could be, is this kind of markup – would that be useful?
Interviewee: I can't give this markup for the translator because they will kill me.
Interviewer: ok, but what about these shows something? Could you tell me how you highlight things?
Interviewee: Well there are different tools that you can use to highlight things, but generally you have to have some kind of input to the software

– maybe it could be something like this. So this is the output that you have right. You don't have the speech? Because this one you did there is manually right?
Interviewer: So, I have just done a sample of them manually.
Interviewee: So, if your output would be like the text, even if the tags come in tags <> but it was highlighted like you did manually here. If they were highlighted, colour coded or something like that, this is something that translators can use because you see the translators need the tags to the first day so then they can follow up what's going on. So in the same way that you needed the full text to highlight, they will need the same text to highlight.
Interviewer: OK, so the challenge with that is that some of these overlap. Some of these are overlapping so here for example: this is finding an anaphora here, and it is finding 136 and 140 but within the same so we have we have we have we have is 136, but then we have to and we have to and we have to is 140 so that's why I was working with a standoff markup because I tried to do it inline first but I kept meeting problems like this or else where, here for example, this has been highlighted as two different ones or this word like repeal here is a blue and a green. So it's an epistrophe and a polyptoton. So marking it up and then, I'm only looking at 4 devices but you could have twelve of them in each.
Interviewee: But maybe are there times that you just want to look into one of those?
Interviewer: Yes, but there's a challenge here with which one do you mark up? How do you choose which one is the right one?
Interviewee: I don't know.
Interviewer: I know!
Interviewee: that's the big question!
Interviewer: And I guess this is all part of the nuance of looking at these.
Interviewee: I don't know if.... Because you're looking into applications what this could be useful for right?
Interviewer: Yeah
Interviewee: You mean what this ouput here? Like it is here – what this could be useful for?
Interviewer: Well, not necessarily. One of the main questions is : is this type of question useful? Would that be of value to have more understanding about these? If so, what would be the ideal format? So you're saying some kind of highlighting?

Interviewee: Yes, for post-editing yes. It would be because translators are humans so they need to see that. A machine can read this, but a human would spend too much time trying to look. Have you talked to anyone who does discourse in machine translation because that is the only thing that comes to my head. I don't know who is working on that though to be honest.... I think [NAME] is working on that discourse for machine translation. But yeah, because I can only tell you from after the machine translation is what I can do with this.
Interviewer: How does your tool, say if you have something like this – how does your tool mark it?
Interviewee: I have one tool that can mark things but it's been a while since I don't do it. I think I include a plain text file with tags where a tool has to colour the words.
Interviewer: It wouldn't be like these tags though?
Interviewee: Maybe we would have the sentence id and then maybe this part, and the element I was. And then. Yeah, maybe something like this.
Interviewer: Cos I guess this is telling it that in sentence 2, the 'I' is the 16th character
Interviewee: Yes, but I don't remember if it's like the tags because it's been like maybe 5-6 years since I used the tool so I don't remember if it was the tags within the text so I had to had the text
Interviewer: tagged already?
Interviewee: yeah, and then I would have here I was and then I would tag this and say colour something or if I have a text on a sentence level and then I have an attribute file on a sentence level and then I would say sentence 1 and then I would have this and then I would have the colour that I want for example, but it could be poss. It could be both. It could be this one. But then you see they would have to be like if I have an attribute file, then I would have to have one per line so for example.
Interviewer: Yeah and then I guess you would have a polysyndeton file or an epanaphora file would you? You'd look at them all separately?
Interviewee: Maybe I could have just one so like if I have a file, the first line of my file was "I welcome the opportunity ..." and to "...the constitution" and then this is the second line right so then in the first line of my attribute file would be 'first line: empty' second line and then I would have sentence id 2, element 'I was' and then I could use this I don't know if that would be necessary or not unless there is repetition of I was more than.
Interviewer: There can be..., it's not with 'I was'

Interviewee: But then again, I would want both 'I was' highlighted no?
Interviewer: Yes
Interviewee: And then I would select a colour but it had to be in the same line to match this. These two files had to be aligned. Sentence 1 here empty, sentence 2 'I was' highlighted, sentence 3 'I was' highlighted, 'to vote in' highlighted. Sentence 4 empty. Sentence 5... you know. That would be something.
Interviewer: OK, what was the name of that tool?
Interviewee: It's called PET. Post-editing tool.
Interviewer: hmmm. I've read something about that.
Interviewee: It's been [UNCLEAR] I don't use it.
Interviewer: Sorry, I'm putting you on the spot.
Interviewee: There are another few ones cos sometimes they have new tools every day but some of them are not very good and some of them you just forget so I don't know.
Interviewer: yes, I'll have a look at it. Again, you know I don't need to actually implement this, it's just more to....
Interviewee: Just know that it is.
Interviewer: Yeah but...
Interviewee: Yeah but like it is something that is very feasible if you just need to discuss it to have like a file like this that shows exactly what has to be highlighted for the translator, that is something that is feasible and it is useful if you are looking into something specific in a text.
Interviewer: Like this, it's.... And it is useful to highlight for post-editors.
Interviewee: Yeah
Interviewer: And I can see how it would be, especially if you have got really long text that you just want to focus in on. And then some of these blue ones, they're done using wordnet and sometimes it comes out as.... Looking at some of them, where I'm not so sure that they actually are cognates. Here's one: it's 'give' and it;s matching it with 'generation'. So, I'm not sure. I need to look at wordnet and see how it's pairing up 'give' and 'generation'. And then there's another one which is even stranger but I guess even if....
Interviewee: What is the blue one?
Interviewer: It's polyptoton. The cognates.
Interviewee: give and generation
Interviewer: And then there's another one. 'Fly' and 'five' if I can find it. I don't know if I'll be able to find it now, I should have marked it. Oh yes, there it is look: 'fly' and 'five'. It's matching this 'five women' and 'fly abroad'. I'm

not sure – I think that's coming from wordnet or wherever they are pulling this from. But I guess if it is highlighting it for a posteditor, then the post editor can say – "No, that's you know….."
Interviewee: Yeah, or they would go "oh my god there is something here I need to pay attention and I don't know what it is"! And then they go like what is it?
Interviewer: Potentially, yeah
Interviewee: Humans!
Interviewer: Humans, yeah. So I need to look into why the tool is doing this, but again, that's not…. My key focus isn't making this work right now anyway.
Interviewee: You know, I think if you already talked to [NAME], from that list that I gave you, I think you should go to [NAME] first because s/he also works with parsing and stuff like that and s/he knows what everyone is doing at all the time. I don't know how they do that. But s/he works with parsing and I'm pretty sure s/he will have many more ideas than me. And with someone that does discourse MT, I'm not sure who's doing that.
REDACTED as not relevant to the discussion)

Interviewee: I have a feeling that discourse MT. I don't know how they implement it but they are trying to get…. Because you know machine translation works at the sentence level so because these features go beyond the sentence level, maybe they do need a corpus that is annotated for that, but then I wouldn't know per se.
Interviewer: And actually someone has done something related but it's not exactly what I'm looking at but they look at metaphor…
The rest of this interview is redacted as it contains sensitive information about ongoing research.

Interview 4
10 December 2018

Interviewer: Is this something that would fit in with your work?
Interviewee: Well, what I work on is, I'm working with researchers who are physicians, [NATIONALITY] physicians that need to publish their research in English. They write their abstracts or their papers in [LANGUAGE] and then we go through Machine Translation and then they do the post-editing.
Interviewer: OK, so they do the post-editing themselves?
Interviewee: uhmhmm. So the way I can see it would fit is for people that – not for my research exactly – but for people who are interested in looking at rhetorical figures in academic speech and that's a whole research area. It's a big one. And it could be interesting to see whether they maintain the same like what they use in their mother tongue and once they have in [UNCLEAR] translation, they keep it. Or they try to keep those voices or they don't. Because they are not native speakers and they don't realise. That's the way I would see it would fit.
Interviewer: Is that? Because I haven't come across literature on that in terms of academic speech? I have found some literature in terms of advertising – maintaining across in advertising.
Interviewee: I don't know if there is anyone. I know that there were people working on that from .... [NAME]. I think it's [NAME] but I'm not sure. From the [UNIVERSITY NAME]. I know that the second, the surname – you know that we have two surnames? The second one is [NAME]. S/he has been doing research on academic. And I know s/he has a big corpus of academic papers and sometimes s/he has, like if there has been a translation, s/he has both. Otherwise s/he has only the Spanish or the English and s/he has even interviews with them like to see how people write their papers. That's the only thing. I know that she and other people from the [UNIVERSITY NAME] have been working so it would be good to see who they are citing. Or who cites them and you would find out. I haven't worked on that so anyway.....
Interviewer: OK, so this wouldn't really fit in with what you work on exactly?
Interviewee: Not really, because what I am looking at is whether or not a non-native speaker of English who is not a professional translator can do post-editing without any training. So it's a completely different [UNCLEAR]
Interviewer: Completely different....

Interviewee: But it could be interesting to see that. Well, I guess.... You have talked to [NAME]. S/he has been working on literary translation. There you would see, because in literary works you would have these kinds of figures. You could see whether or not machine translation system was working.
Interviewer: Yeah, absolutely. So one of the things that was coming out of the interviews that I've done so far is that this could be useful in terms of highlighting in a post-editing system where you might highlight something in the source language and to keep an eye on it in the target language.
Interviewee: That would be good for a professional translator. Back in the day before I even did my PhD, I was working in a translation company as a translator. Once we had like, one of our clients was the regional government of my hometown. And they would translate the speeches from Obama ...
Interviewer: Oh really?
Interviewee: Yeah, and other politicians around the world because that was the way for the politicians in [COUNTRY] to get inspired. So we would get this as a task, we were translating the speeches from other politicians from English into [LANGUAGE] because then their office would look at what they have been talking about and what rhetorical speeches, eh figures they were using. So that later on....
Interviewer: And was that something that they were actively looking at then?
Interviewee: I am assuming that's what their office was looking after. I mean we were just hired as for translation but I am sure that was... eh like if a politician is known to have good speeches that's usually.... And we have like in [COUNTRY], as in every country, you have very good speakers and very bad speakers. So I guess they were trying to get better at public speaking and convincing the citizens to vote for them.
Interviewer: And as translators, were you aware of these and of carrying these over?
Interviewee: We didn't really look into that.
Interviewer: Yeah, cos some of them are quite subtle.
Interviewee: I wasn't doing the translations, I was just like operations manager so I was just assigning tasks. I was not in charge of even having a look at it. But I would assume that the translators we had would have a look at that. We also had.... We were working with marketing companies. In this case, it was [LANGUAGE] into English and in that case, we had like slogans and things like that. I know the translators were trying to do the same like. Okay, there are these sounds that are repeated and I have to try to repeat, eh

a similar sound. And sometimes we had problems because of the culture. And sometimes they were like, OK I get the point and in [NATIONALITY] it sounds brilliant, but this does not translate well in English and they would try to come up with something that would be also like... if there was a repetition, a different type of repetition and had to argue with the client – OK this is not a direct translation but it achieves the same goal which is what you're after when you are doing marketing
Interviewer: ya, just selling. Yeah, because I have seen some research into translating these kinds of devices and how they don't always ....
Interviewee: they don't always correspond
Interviewer: No, a polyptoton in English might not work in Russian for example.
Interviewee: And also, I would assume that the, what was the name? the polyptoton might not always be the same as well.
Interviewer: Yeah, depending on....
Interviewee: So those are the things that I can come up with.
Interviewer: OK. Yeah. And do you think that, so, do you think there is a value in this type of ...?
Interviewee: I think it's... I mean everything can be useful for different purposes right? It all depends who you are talking to. The fact that it is not useful for my research because I am focusing on what type of corrections they are able to make doesn't mean that it's not relevant. And as I said, there is a whole research area in writing research that probably is looking after those things. There are so many things like why, I am sure there are many researchers now looking at influencers, the way they talk and why are they influencers. Maybe they are using this kind of things.
Interviewer: Possibly....
Interviewee: But what is also interesting from a psychological point of view is whether they are consciously choosing to use these rhetorical figures or they just like use them.
Interviewer: Yeah. And I think that's something that as you go through these, it's hard to you know, we don't know which ones are used deliberately and which ones are just used because we use them all the time without knowing anyway.
Interviewee: When you are repeating a lot like the politicians, I guess that that one is, if you choose to use that one, it's because you want to stress that one. The same way as when you are repeating the beginning or the end. I think that's more conscious, whereas the polyptoton, I would assume, that's

my hypothesis that it's more subconscious. That you are trying to do the.... And I would imagine that you would have more of these things in at least [LANGUAGE]. We are very very very verbose.
Interviewer: Yeah.
Interviewee: So I guess a way of actually making, especially in academic writing, but what I have seen of repeating the same idea once, because we do circles until we get to the point, would be to use different ways of saying the same thing. So you would do ???? more of these probably.
Interviewer: Yes, that's interesting. And, you know when learning how to write in [LANGUAGE], are you taught these techniques or....?
Interviewee: Well, I remember when I started doing my PhD that I was told to be more like.... Like for me it was an act of pro-action to actually say things in a short sentence. [LANGUAGE] academic writing is so different, right? So when I was writing in English, I was using strategies I would use in [LANGUAGE]. And then my supervisor, I did my PhD in [COUNTRY].
Interviewer: OK
Interviewee: My supervisor was from [COUNTRY] so, s/he was like "You make too long sentences, you have to go to the point" and that's how I learned actually that I was very repetitive, without noticing it. It would be maybe interesting to see if you are writing maybe in a different language – I was writing in English, but I was still using the same rhetorical figures that I would use in [LANGUAGE].
Interviewer: Yeah, OK
Interviewee: So there might be a pattern. Maybe there is a way of like correlating whether the fact that you are like based in [COUNTRY] and are not having an international environment influences the way you are writing.
Interviewer: And I'm sure it does. OK, I'll definitely look into this work and....
Interviewee: I don't know if there are other people working on in the different languages that would be more prone to one or the other.
Interviewer: It's something that I haven't come across but then I'm not ....
Interviewee: Well, that's not the goal of your research.
Interviewer: Exactly. So I'm not even looking at how these might be translated, it's more how this might be implemented in a translation system. But yeah, I'll definitely have a look at this work to see. Let me just have a look at my questions. I'm not sure. If it's not directly relating to your own work if it's ehm..... Could I just ask you about the TEI that you use?
Interviewee: It was back in 2011, so it was ....

Interviewer: OK, so you're not using it currently?
Interviewee: I just, the corpus of my PhD, I released it publicly and then I encoded it in TEI so that anyone could use it. It was both TEI and Translation Memory. I encoded it in two different ways. And I was looking at the translation of compound words from [LANGUAGE] into [LANGUAGE] – I don't know if you know any [LANGUAGE], but maybe....
Interviewer: Yeah, I did [LANGUAGE] and [LANGUAGE] ....
Interviewee: but they [UNCLEAR] the words right? In English, you would just split the compound and pretty much translate word by word but in [LANGUAGE] you have to reverse the order because the head of the compound is always at the very end at the right, and then in [LANGUAGE] the nucleus of the nominal phrase has to be on the beginning. And then you have to add prepositions and determiners because otherwise it doesn't make any sense so take the complements of a noun come in a different.... Like the syntactical structure is completely different. So I was trying to annotate that in TEI but it was a disaster. The moment you try to annotate those kinds of complicated things, it gets complicated.
Interviewer: That's what I found with these as well because I initially started out looking at TEI because that's what I've used before. And then I was trying to do inline markup and it just .... I kept running into difficulties with things where, I don't know if there's actually a great example of it here.... Even things like this one – it was finding these were two different anaphora or epanaphora so: we have, we have, we have, we have, we have but then it was also finding "we have to" "we have to" and "we have to" but then .... So the inline markup broke down very quickly, so that's why following this one..... And the other mark up that you used – was that TMX? Translation Memory
Interviewee: Exchange, yeah. And then I have been involved in, so it has nothing to do with translation, and I don't know, do you know what multiword expressions are? Have you heard of them?
Interviewer: Yes
Interviewee: I have been involved in a very large initiative. You have talked to [NAME]? S/he has also been participating in the shared tasks, trying to annotate multi-word expressions and we have been using a completely different system cause it was like nothing else was working. Cause we needed to annotate things that were even discontinuous like in [LANGUAGE] a very that has a particle at the end of the sentence. Things like that. That creates a lot of problems when you try to put them together and then we also

had cases like what you have here with the "have to". Sometimes we had a multiword expression that is embedded into another multiword expression and we need to be able to annotate both. So s/he created our own standard which is basically an XML but I don't remember.
Interviewer: Ya, S/he explained it to me before. S/he showed me it.
Interviewee: S/he know better than I because I have been working on the interface – I haven't worked on the files – that's a computer scientist – they were generating them automatically. But I know [NAME] has worked a lot with them because s/he was doing some analysis on the data.
Interviewer: Ya, S/he did explain that to me before. And what was the work that you were doing on the interface then?
Interviewee: Annotations. S/he created an annotation interface for us. So the linguists could go and just signal which are the words that you want to annotate. And then you just select.
Interviewer: OK.
Interviewee: It was very straightforward. That's another possible application.
Interviewer: Ya, I was just thinking that, because...
Interviewee: It's not translation again, it's multiword expressions. You have things like if you have to vote in, although it's different in that campaign in... 1993.
Interviewer: Oh so in a multiword expression, what follows is ....
Interviewee: Well, in a multiword expression, the idea is that you need all the elements of the multiword expression in order to have meaning. Otherwise it's ungrammatical. Like you have "to take a shower" – you cannot "have a shower" in English or you cannot "make a picture" in English, you "take a picture". Whereas in [LANGUAGE] you "make a picture". That's how we say it. And that goes together. You cannot change the verb because it would have a different meaning. In some cases, if you change the verb then it has a completely different meaning. In English you would have all the phrasal verbs and prepositional verbs aswel. You could also have idioms like "it's raining cats and dogs" and then the more obscure ones are the ones like together they mean something that is not the sum of the individual meanings of the components. And last year we included things like when the verb is asking for a specific preposition and then you could have that the same verb that asks for different prepositions depending on the meaning.
Interviewer: OK, uh huh
Interviewee: So this would not be what we call a multiword expression in

that research area.
Interviewer: yeah, yeah. OK,
Interviewee: But it's like "to cast an opinion" or "to cast a view", you would have to apply the, like there are tests to see if in this case cast is being used as a light verb and what is giving actually the full meaning is the noun that goes with it. But that's a completely different thing, unrelated to your research.
Interviewer: It is! Hahaha! But it is still interesting even in terms of the nuance of this type of language and I guess for multiword expressions aswel.
Interviewee: The thing is if you have this, that to cast an opinion and to cast a view are two light verb constructions. The moment you are translating them it might be that you are not having cast anymore, that in a different language you are having two different verbs. And then you cannot maintain this repetition anymore unless you change completely. And then in that case, maybe the translator chooses to do something as when I was studying translation, they were actually telling us if you see that kind of thing and you cannot repeat it where it appears, what you have to do is try to find ...
Interviewer: somewhere else to include it?
Interviewee: Somewhere else to include it.
Interviewer: OK. Even if it's down here somewhere?
Interviewee: Even with jokes, they would say if there is a wordplay, and you cannot repeat it in I was doing English into [LANGUAGE] right? If you cannot repeat it in [LANGUAGE], because it doesn't work, you have to look for a place in the text where you could do the wordplay that it would fit because at the end of the day, you want to have the same experience in the reader. If the original reader was having a laugh there, you want the reader of the translation to have a laugh somewhere near by that area. So there are translation theories and models that suggest that.
Interviewer: Yeah, I must look at those as well so. OK. Well I think that's kind of covered it, because some of my questions are "how would YOU implement this in your work?" whereas if....
Interviewee: Well, I guess for literary translation, this would be very useful. Because they don't necessarily use machine translation, they use translators. But it could be a way of knowing when they have to actually look after these things because most of the time, when you are a translator, you have to translate 3000 words in a day. That's your target. And you are going at such a speed that you don't have time to have a look at these things.
Interviewer: Yeah, and to try to find something that might be spread across a whole paragraph or....

Interviewee: Oh yeah, if it is far away then it's even... Because you are going sentence by sentence. So if you see, if you have these things highlighted in the original text somehow, so that it's like be careful because this is happening. For literary translators, I think that would be great. Or for anyone working with marketing campaigns or .... Probably not in technical documents.
Interviewer: No! And I think that's what's coming out of you know from talking to people is that that is how this might be, you know because I spoke to people who would see these as features and might try to build them into a system but they were saying with neural machine translation now it's very difficult to ....
Interviewee: Yeah, but don't think only. I wouldn't think only machine translation, I would also think of the translators. Or even if they use machine translation, when they were doing the post-editing. The machine translation might be brilliant, but if it's not repeating the same things, then your job as a post-editor might be to actually pick those things and see them.
Interviewer: Absolutely.

Bibliography

About WordNet (2010). URL: https://wordnet.princeton.edu/ (visited on 02/01/2019).

Allen, Jeffrey (2001). "Postediting: An Integrated Part of a Translation Software Program". In: *Language International* 13.2, pp. 26–29.

An Overview of the Human Genome Project (2016). URL: https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/ (visited on 01/17/2019).

Anderson, Steve and Tara McPherson (2011). "Engaging Digital Scholarship: Thoughts on Evaluating Multimedia Scholarship". In: *Profession* 2011.1, pp. 136–151.

Arnold, Doug (2003). "Why Translation Is Difficult for Computers". In: *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins Publishing, pp. 119–42.

Atkinson, J. Maxwell (2004). *Lend Me Your Ears: All You Need to Know about Making Speeches and Presentations*. London: Vermilion. 376 pp.

Aziz, Wilker, Sheila Castilho, and Lucia Specia (2012). "PET: A Tool for Post-Editing and Assessing Machine Translation." In: *LREC*, pp. 3982–3987.

Bailis, Stanley (2001). "Contending with Complexity: A Response to William H. Newell's" A Theory of Interdisciplinary Studies"". In: *Issues in Interdisciplinary Studies*.

Baker, Paul (2006). *Using Corpora in Discourse Analysis*. A&C Black.

Bang, Henrik P. (2009). "'Yes We Can': Identity Politics and Project Politics for a Late-Modern World". In: *Urban Research & Practice* 2.2, pp. 117–137.

Barković, Dražen (2010). "Challenges of Interdisciplinary Research". In: *Interdisciplinary Management Research* 6, pp. 951–960.

Bell, Judith (2014). *Doing Your Research Project: A Guide for First-Time Researchers*. McGraw-Hill Education (UK).

Bentham, Transcribe. *About Us - Transcribe Bentham*. URL: http://blogs.ucl.ac.uk/transcribe-bentham/about/ (visited on 12/17/2018).

Berez-Kroeker, Andrea L. et al. (2018). "The Austin Principles of Data Citation in Linguistics". In: URL: https://site.uit.no/linguisticsdatacitation/austinprinciples/ (visited on 07/30/2018).

Bhavsar, Mit (2017). "Multidisciplinary Research: Pros and Cons". In: *Nature Jobs*. URL: http://blogs.nature.com/naturejobs/2017/09/11/multidisciplinary-research-pros-and-cons/.

Biancani, Susan et al. (2018). "Superstars in the Making? The Broad Effects of Interdisciplinary Centers". In: *Research Policy* 47.3, pp. 543–557.

Blackmore, Karen L. and Keith V. Nesbitt (2008). "Identifying Risks for Cross-Disciplinary Higher Degree Research Students". In: *Proceedings of the Tenth Conference on Australasian Computing Education-Volume 78*. Australian Computer Society, Inc., pp. 43–52.

Bungay, Stephen. "His Speeches: How Churchill Did It". In: *The International Churchill Society*. URL: https://winstonchurchill.org/resources/speeches/speeches-about-winston-churchill/his-speeches-how-churchill-did-it/ (visited on 11/19/2018).

Burton, Gideon O. *Silva Rhetoricae: The Forest of Rhetoric*. URL: http://rhetoric.byu.edu/ (visited on 11/07/2016).

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006). "Re-Evaluation the Role of Bleu in Machine Translation Research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Campbell, Margaret C. and Kevin Lane Keller (2003). "Brand Familiarity and Advertising Repetition Effects". In: *Journal of consumer research* 30.2, pp. 292–304.

Castilho, Sheila et al. (2017). "Is Neural Machine Translation the New State of the Art?" In: *The Prague Bulletin of Mathematical Linguistics* 108.1, pp. 109–120.

Causer, Tim et al. (2018). "'Making Such Bargain': Transcribe Bentham and the Quality and Cost-Effectiveness of Crowdsourced Transcription". In: *Digital Scholarship in the Humanities*.

Cavanagh, Sheila (2012). "Living in a Digital World: Rethinking Peer Review, Collaboration, and Open Access". In: *ABO: Interactive Journal for Women in the Arts, 1640-1830* 2.1, p. 15.

Chaffey, Dave and Gareth White (2010). *Business Information Management: Improving Performance Using Information Systems*. Pearson Education.

Charteris-Black, Jonathan (2011). *Politicians and Rhetoric: The Persuasive Power of Metaphor*. 2nd ed. Basingstoke: Palgrave Macmillan. 370 pp. ISBN: 978-0-230-25164-9 0-230-25164-1 978-0-230-25165-6 0-230-25165-X.

Choi, Bernard CK and Anita WP Pak (2006). "Multidisciplinarity, Interdisciplinarity and Transdisciplinarity in Health Research, Services, Education

and Policy: 1. Definitions, Objectives, and Evidence of Effectiveness". In: *Clinical and investigative medicine* 29.6, p. 351.

Chozick, Amy (2015). "Hillary Clinton's Beijing Speech on Women Resonates 20 Years Later". In: *The New York Times. Politics*. ISSN: 0362-4331. URL: https://www.nytimes.com/politics/first-draft/2015/09/05/20-years-later-hillary-clintons-beijing-speech-on-women-resonates/ (visited on 01/26/2019).

Cohen, Fernand (2018). "Cultural Machine Translation–Challenges and Solutions". In: MedPRAI 2018. Rabat Morocco. URL: https://medprai2018.sciencesconf.org/.

Conlan, Owen et al. (2014). "Revolutionary Entities: Turning Data into Knowledge to Drive Personalized Exploration of The Irish Rising of 1916". In: *Big Data (Big Data), 2014 IEEE International Conference On*. IEEE, pp. 32–38.

Cowan, Robin and Dominique Foray (1997). "The Economics of Codification and the Diffusion of Knowledge". In: *Industrial and corporate change* 6.3, pp. 595–622.

Creswell, John W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Fourth edition, international student edition. Los Angeles, Calif: SAGE. 273 pp. ISBN: 978-1-4522-7461-4 978-1-4522-7460-7.

Crines, Andrew S. (2016). "The Rhetoric of the EU Referendum Campaign". In: *EU Referendum Analysis 2016*. URL: https://www.referendumanalysis.eu/eu-referendum-analysis-2016/section-5-campaign-and-political-communication/the-rhetoric-of-the-eu-referendum-campaign/ (visited on 07/25/2018).

CULTURA (Cultivating Understanding and Research through Adaptivity). URL: http://www.cultura-strep.eu/cultura-vision (visited on 01/12/2019).

David, Maya Khemlani (2014). "Language, Power and Manipulation: The Use of Rhetoric in Maintaining Political Influence". In: *Frontiers of Language and Teaching* 5. URL: http://bit.ly/MayaKhemlaniDavid (visited on 11/02/2016).

Delgado, Ana and Heidrun Åm (2018). "Experiments in Interdisciplinarity: Responsible Research and Innovation and the Public Good". In: *PLoS biology* 16.3, e2003921.

Della Chiesa, Bruno, Vanessa Christoph, and Christina Hinton (2009). "How Many Brains Does It Take to Build a New Light: Knowledge Management Challenges of a Transdisciplinary Project". In: *Mind, Brain, and Education* 3.1, pp. 17–26.

DeRose, Steven J. (2004). "Markup Overlap: A Review and a Horse." In: *Extreme Markup Languages®*.

Di Cresce, Rachel and Julia King (2017). "Developing Collaborative Best Practices for Digital Humanities Data Collection: A Case Study". In: *College & Undergraduate Libraries* 24.2-4, pp. 226–237.

Dorr, Bonnie J., Pamela W. Jordan, and John W. Benoit (1999). "A Survey of Current Paradigms in Machine Translation". In: *Advances in Computers*. Vol. 49. Elsevier, pp. 1–68.

Dubremetz, Marie (2017). *Detecting Rhetorical Figures Based on Repetition of Words: Chiasmus, Epanaphora, Epiphora*. Acta Universitatis Upsaliensis.

Dubremetz, Marie and Joakim Nivre (2018). "Rhetorical Figure Detection: Chiasmus, Epanaphora, Epiphora". In: *Frontiers in Digital Humanities* 5, p. 10.

Dykes, Thomas H., Paul A. Rodgers, and Michael Smyth (2009). "Towards a New Disciplinary Framework for Contemporary Creative Design Practice". In: *CoDesign* 5.2, pp. 99–116.

Edmond, Jennifer, Naveen Bagalkot, and Alex O'Connor (2016). *Toward a Deeper Understanding of the Scientific Method of the Humanist*.

Fahnestock, Jeanne (1999). *Rhetorical Figures in Science*. Oxford University Press.

Fitzpatrick, Kathleen (2012). "The Humanities, Done Digitally". In: *Debates in the Digital Humanities*. Ed. by Matthew K. Gold. Minneapolis: University of Minnesota Press. URL: http://dhdebates.gc.cuny.edu/debates/text/30 (visited on 07/31/2018).

Flanders, Julia (2013). "The Productive Unease of 21st-Century Digital Scholarship". In: *Defining Digital Humanities: A Reader*, pp. 205–218.

Forsyth, Mark (2014). *The Elements of Eloquence: How To Turn the Perfect English Phrase*. Icon Books Ltd. 224 pp. ISBN: 978-1-84831-733-8.

Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019). URL: https://www.idiap.ch/workshop/DiscoMT (visited on 11/19/2018).

Franks, Daniel et al. (2007). "Interdisciplinary Foundations: Reflecting on Interdisciplinarity and Three Decades of Teaching and Research at Griffith University, Australia". In: *Studies in Higher Education* 32.2, pp. 167–185.

Garcia, Eva Martínez et al. (2017). "Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation". In: *The Prague Bulletin of Mathematical Linguistics* 108.1, pp. 85–96.

Gawryjolek, Jakub Jan (2009). *Automated Annotation and Visualization of Rhetorical Figures*. University of Waterloo.

Girard, John and John Girard (2015). "Defining Knowledge Management: Toward an Applied Compendium". In: *Online Journal of Applied Knowledge Management* 3.1, pp. 1–20.

Godin, Benoit (2006). "The Knowledge-Based Economy: Conceptual Framework or Buzzword?" In: *The Journal of technology transfer* 31.1, pp. 17–30.

Goodman, Robert et al. (2007). "The Use of Stylometry for Email Author Identification: A Feasibility Study". In: *Proc. Student/Faculty Research Day, CSIS, Pace University, White Plains, NY*, pp. 1–7.

Gordon, Robert M. et al. (2014). "A Transdisciplinary Team Approach to Pain Management in Inpatient Health Care Settings". In: *Pain Management Nursing* 15.1, pp. 426–435.

Green, Spence, Jeffrey Heer, and Christopher D. Manning (2013). "The Efficacy of Human Post-Editing for Language Translation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 439–448.

Griffin, Gabriele and Matt Hayler (2016). *Research Methods for Reading Digital Data in the Digital Humanities*. Edinburgh University Press. 223 pp. ISBN: 978-1-4744-0962-9.

Griffin, Gabriele and Matt Steven Hayler (2018). "Collaboration in Digital Humanities Research – Persisting Silences". In: *Digital Humanities Quarterly* 012.1. ISSN: 1938-4122. URL: http://digitalhumanities.org/dhq/vol/12/1/000351/000351.html (visited on 10/17/2018).

Group, Stanford NLP. *Tokenization*. URL: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html (visited on 02/22/2019).

Hall, Hagey, Davis Centre, and University of Waterloo Canada. *Computational Rhetoric Workshop*. URL: https://uwaterloo.ca/arts/events/computational-rhetoric-workshop (visited on 11/15/2018).

Hardmeier, Christian (2012). "Discourse in Statistical Machine Translation. a Survey and a Case Study". In: *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 11.

Harris, Randy and Chrysanne DiMarco (2009). "Constructing a Rhetorical Figuration Ontology". In: *Persuasive Technology and Digital Behaviour Intervention Symposium*. Citeseer, pp. 47–52.

Harris, Randy Allen and Chrysanne Di Marco (2017). "Rhetorical Figures, Arguments, Computation". In: *Argument & Computation* 8.3, pp. 211–231.

Harris, Randy Allen et al. (2018). "An Annotation Scheme for Rhetorical Figures". In: *Argument & Computation* (Preprint), pp. 1–21.

Hayler, Matt and Gabriele Griffin (2016). *Research Methods for Creating and Curating Data in the Digital Humanities*. Edinburgh University Press. 298 pp. ISBN: 978-1-4744-0967-4.

Hayles, N. Katherine (2012). "How We Think: Transforming Power and Digital Technologies". In: *Understanding Digital Humanities*. Springer, pp. 42–66.

Hearne, Mary and Andy Way (2011). "Statistical Machine Translation: A Guide for Linguists and Translators". In: *Language and Linguistics Compass* 5.5, pp. 205–226.

Hill, Teresa Garrett et al. (2010). "Evaluation of Cancer 101: An Educational Program for Native Settings". In: *Journal of cancer education : the official journal of the American Association for Cancer Education* 25.3, pp. 329–336. ISSN: 0885-8195. DOI: 10.1007/s13187-010-0046-5. pmid: 20146041. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935509/ (visited on 11/16/2018).

Holbrook, J. Britt (2013). "What Is Interdisciplinary Communication? Reflections on the Very Idea of Disciplinary Integration". In: *Synthese* 190.11, pp. 1865–1879.

Hromada, Daniel Devatman (2011). "Initial Experiments with Multilingual Extraction of Rhetoric Figures by Means of PERL-Compatible Regular Expressions". In: *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pp. 85–90.

Huutoniemi, Katri (2010). *Evaluating Interdisciplinary Research*. Vol. 10. Oxford University Press Oxford.

Jacobs, Jerry A. and Scott Frickel (2009). "Interdisciplinarity: A Critical Assessment". In: *Annual review of Sociology* 35.

Jarvenpaa, Sirkka L. and D. Sandy Staples (2001). "Exploring Perceptions of Organizational Ownership of Information and Expertise". In: *Journal of Management Information Systems* 18.1, pp. 151–183.

Jashapara, Ashok (2004). *Knowledge Management: An Integrated Approach*. Pearson Education.

Java, James (2015). *Characterization of Prose by Rhetorical Structure for Machine Learning Classification*. Florida, USA: Nova Southeastern University. URL: http://nsuworks.nova.edu/gscis_etd/347.

Johnson, Björn, Edward Lorenz, and Bengt-\AAke Lundvall (2002). "Why All This Fuss about Codified and Tacit Knowledge?" In: *Industrial and corporate change* 11.2, pp. 245–262.

Johnson, Melissa (2015). "Women's Rights Are Human Rights". URL: https://prezi.com/nokvkg57pbzx/womens-rights-are-human-rights/ (visited on 01/26/2019).

Jones, Casey (2010). "Interdisciplinary Approach-Advantages, Disadvantages, and the Future Benefits of Interdisciplinary Studies". In: *Essai* 7.1, p. 26.

Kakutani, Michiko (2015). "Obama's Eulogy, Which Found Its Place in History". In: *The New York Times. Arts*. ISSN: 0362-4331. URL: https://www.nytimes.com/2015/07/04/arts/obamas-eulogy-which-found-its-place-in-history.html (visited on 01/26/2019).

Katan, David (2016). "Translation at the Cross-Roads: Time for the Transcreational Turn?" In: *Perspectives* 24.3, pp. 365–381.

Klein, Julie T. (2008). "Evaluation of Interdisciplinary and Transdisciplinary Research: A Literature Review". In: *American journal of preventive medicine* 35.2, S116–S123.

Klein, Julie Thompson (2010). "A Taxonomy of Interdisciplinarity". In: *The Oxford Handbook of Interdisciplinarity*. Vol. 15, pp. 15–30.

Klein, Julie Thompson and William H. Newell (1997). "Advancing Interdisciplinary Studies". In: *Handbook of the undergraduate curriculum: A comprehensive guide to purposes, structures, practices, and change*, pp. 393–415.

Klein, Julie Thompson, Jay Wentworth, and David Sebberson (2001). "Interdisciplinarity and the Prospect of Complexity: The Tests of Theory". In: *Issues in Interdisciplinary Studies*.

Koehn, Philipp. *The State of Neural Machine Translation (NMT)*. URL: https://omniscien.com/state-neural-machine-translation-nmt/ (visited on 02/18/2019).

— (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *MT Summit*. Vol. 5, pp. 79–86.

Koppel, Moshe and Jonathan Schler (2003). "Exploiting Stylistic Idiosyncrasies for Authorship Attribution". In: *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Vol. 69, pp. 72–80.

Larivière, Vincent and Yves Gingras (2014). "10 Measuring Interdisciplinarity". In: *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, p. 187.

Lawrence, John, Jacky Visser, and Chris Reed (2017). "Harnessing Rhetorical Figures for Argument Mining". In: *Argument & Computation* 8.3, pp. 289–310.

Leahey, Erin, Christine M. Beckman, and Taryn L. Stanko (2017). "Prominent but Less Productive: The Impact of Interdisciplinarity on Scientists' Research". In: *Administrative Science Quarterly* 62.1, pp. 105–139.

Leahy, Pat. "Referendum Tracker". In: *The Irish Times*. URL: https://www.irishtimes.com/news/politics/referendum-tracker (visited on 03/17/2018).

Leavy, Susan, Emilie Pine, and Mark T. Keane (2017). "Mining the Cultural Memory of Irish Industrial Schools Using Word Embedding and Text Classification". In: *DH 2017*.

— (2018). "Industrial Memories: Exploring the Findings of Government Inquiries with Neural Word Embedding and Machine Learning". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Ledford, Heidi (2015). "How to Solve the World's Biggest Problems". In: *Nature News* 525.7569, p. 308.

Lee, Ming-Chang (2010). "Knowledge-Based New Product Development through Knowledge Transfer and Knowledge Innovation". In: *Innovation through Knowledge Transfer*. Springer, pp. 303–320.

Leech, Geoffrey (2005). *Adding Linguistic Annotation.*

Libovicky, Jindřich and Bruno Cartoni (2018). *Machine Translation Evaluation beyond the Sentence Level*.

LitCharts. URL: https://www.litcharts.com/literary-devices-and-terms/epistrophe (visited on 01/26/2019).

LitCharts. URL: https://www.litcharts.com/literary-devices-and-terms/polyptoton (visited on 01/26/2019).

LitCharts. URL: https://www.litcharts.com/literary-devices-and-terms/polysyndeton (visited on 01/26/2019).

Lopez, Mark Hugo and Paul Taylor (2009). *Dissecting the 2008 Electorate: Most Diverse in US History*. Pew Hispanic Center Washington, DC.

Lundvall, Bengt-äke and Björn Johnson (1994). "The Learning Economy". In: *Journal of industry studies* 1.2, pp. 23–42.

MacLeod, Miles (2018). "What Makes Interdisciplinarity Difficult? Some Consequences of Domain Specificity in Interdisciplinary Practice". In: *Synthese* 195.2, pp. 697–720.

Maps, Deep. *About the Project*. URL: http://www.deepmapscork.ie/about/about-the-project/ (visited on 02/25/2019).

— *Home*. URL: http://www.deepmapscork.ie/ (visited on 02/25/2019).

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". In:

Martin, Jonathan (2016). "Donald Trump's Barrage of Heated Rhetoric Has Little Precedent". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2016/10/15/us/politics/trump-speech-highlights.html (visited on 11/24/2016).

Mattern, Shannon Christine (2012). "Evaluating Multimodal Work, Revisited". In: *Words in Space*.

McCarty, Willard and Marilyn Deegan (2016). "Digital Humanities in the Age of the Internet: Reaching out to Other Communities". In: *Collaborative Research in the Digital Humanities*. Routledge, pp. 93–104.

McKeon, Michael (1994). "The Origins of Interdisciplinary Studies". In: *Eighteenth-Century Studies* 28.1, pp. 17–28. ISSN: 0013-2586. DOI: 10.2307/2739220. URL: https://www.jstor.org/stable/2739220 (visited on 01/17/2019).

McMurtry, A. et al. (2012). "Making Interdisciplinary Collaboration Work: Key Ideas, a Case Study and Lessons Learned". In: *Alberta Journal of Educational Research* 58.3, pp. 461–473.

Merino, Miguel Bernal (2006). "On the Translation of Video Games". In: *JoSTrans: The Journal of Specialized Translation* 6, p. 29.

Moreno, María del Carmen Calatrava, Petra Kynčlová, and Hannes Werthner (2016). "A Multiple-Perspective Analysis of Doctoral Interdisciplinarity". In: *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, pp. 1–11.

Murphy, Rachel and Orla-Peach Power (2017). "Exploring the Digital Project Lifecycle with Deep Maps: West Cork Coastal Cultures". URL: https://ucc.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=56bd595e-3f78-4f35-b0cb-1970f3e54071 (visited on 12/18/2018).

Nelson, Richard R. (2000). "Knowledge and Innovation Systems". In: *Knowledge Management in the Learning society*, pp. 115–124.

Newell, William H., Jay Wentworth, and David Sebberson (2001). "A Theory of Interdisciplinary Studies". In: *Issues in Interdisciplinary Studies*.

Nowviskie, Bethany (2012). "Evaluating Collaborative Digital Scholarship (or, Where Credit Is Due)". In: *Jounal of Digital Humanities* 1.4, pp. 16–30.

O'Brien, Sharon and Gabriela Saldanha (2014). *Research Methodologies in Translation Studies*. Routledge.

Odacıoğlu, Mehmet Cem and Şaban Köktürk (2015). "From Interdisciplinarity to Transdisciplinarity in Translation Studies in the Context of Technological Tools & Localization Industry". In: *International Journal of Comparative Literature and Translation Studies* 3.3, pp. 14–19.

OECD (1996). *OECD Science, Technology and Industry Outlook 1996: The Knowledge-Based Economy*. Paris: OECD Publishing.

— (2000). *Knowledge Management in the Learning Society*. OECD Paris.

Oireachtas, Houses of the (2017). *Report of the Joint Committee on the Eighth Amendment of the Constitution*. Dublin, Ireland: Houses of the Oireachtas. URL: https://www.oireachtas.ie/parliament/media/committees/eighthamendmentoftheconstitution/Report-of-the-Joint-Committee-on-the-Eighth-Amendment-web-version.pdf.

— (2018). *More about the Service – Houses of the Oireachtas*. URL: https://www.oireachtas.ie/en/how-parliament-is-run/houses-of-the-oireachtas-service/more-about-the-service (visited on 01/28/2019).

O'Reilly, Cliff and Shamima Paurobally (2010). "Lassoing Rhetoric with OWL and SWRL". In: *Unpublished MSc dissertation*. URL: http://bit.ly/LassoingRhetoric.

O'Brien, Sharon and Michel Simard (2014). "Introduction to Special Issue on Post-Editing". In: *Machine Translation* 28.3-4, pp. 159–164.

Papineni, Kishore et al. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 311–318.

Parkinson, R. B. (2005). *The Rosetta Stone*. In collab. with British Museum. British Museum Objects in Focus. London: British Museum. 64 pp. ISBN: 978-0-7141-5021-5.

Patton, Michael Quinn (1990). *Qualitative Evaluation and Research Methods*. SAGE Publications, inc.

Pedersen, Daniel (2014). "Exploring the Concept of Transcreation–Transcreation as 'More than Translation'". In: *Cultus: The Journal of intercultural mediation and communication* 7, pp. 57–71.

Pedersen, David Budtz (2016). "Integrating Social Sciences and Humanities in Interdisciplinary Research". In: *Palgrave Communications* 2, p. 16036.

Peldszus, Andreas and Manfred Stede (2013). "From Argument Diagrams to Argumentation Mining in Texts: A Survey". In: *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7.1, pp. 1–31.

Pine, Emilie, Susan Leavy, and Mark T. Keane (2017). "Re-Reading the Ryan Report: Witnessing via and Close and Distant Reading". In: *Éire-Ireland* 52.1, pp. 198–215.

Piotrowski, Michael (2012). *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers. 159 pp.

PMI, Project Management Institute (2015). "Capturing the Value of Project Management Through Knowledge Transfer". In: URL: https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/capture-value-knowledge-transfer.pdf (visited on 11/03/2018).

Pray, Leslie (2002). "Interdisciplinarity in Science and Engineering: Academia in Transition". In: *Science Career Magazine*.

Páez, Mariela and Lillie R. Albert (2012). "Cultural Consciousness". In: *Encyclopedia of Diversity in Education*. 4 vols. Thousand Oaks: SAGE Publications, Inc., pp. 510–510. DOI: 10.4135/9781452218533. URL: http://sk.sagepub.com/reference/diversityineducation/n160.xml (visited on 11/13/2018).

Quoc, V. Le and Mike Schuster (2016). *A Neural Network for Machine Translation, at Production Scale*. URL: http://ai.googleblog.com/2016/09/a-neural-network-for-machine.html (visited on 08/08/2018).

Rawlings, Craig M. et al. (2015). "Streams of Thought: Knowledge Flows and Intellectual Cohesion in a Multidisciplinary Era". In: *Social Forces* 93.4, pp. 1687–1722.

Rhoten, Diana (2004). "Interdisciplinary Research: Trend or Transition". In: *Items and Issues* 5.1-2, pp. 6–11.

Rike, Sissel Marie (2013). *Bilingual Corporate Websites-from Translation to Transcreation?*

Rivera, Y. M. et al. (2016). "When a Common Language Is Not Enough: Transcreating Cancer 101 for Communities in Puerto Rico". In: *Journal of Cancer Education* 31.4, pp. 776–783.

Robertson, David W., Douglas K. Martin, and Peter A. Singer (2003). "Interdisciplinary Research: Putting the Methods under the Microscope". In: *BMC Medical Research Methodology* 3.1, p. 20.

Rockwell, Geoffrey (2012). *Short Guide To Evaluation Of Digital Work*. URL: http://journalofdigitalhumanities.org/1-4/short-guide-to-evaluation-of-digital-work-by-geoffrey-rockwell/ (visited on 01/17/2019).

Rodríguez, Ayuso IR (2012). "Puerto Rico Chronic Diseases Report 2012". In: *Puerto Rico Health Department*.

Roukos, Salim, David Graff, and Dan Melamed (1995). *Hansard French/English LDC95T20*. Philadelphia: Linguistic DAta Consortium. URL: https://catalog.ldc.upenn.edu/LDC95T20 (visited on 08/13/2018).

Rowley, Jennifer (2007). "The Wisdom Hierarchy: Representations of the DIKW Hierarchy". In: *Journal of information science* 33.2, pp. 163–180.

Rylance, Rick (2015). "Grant Giving: Global Funders to Focus on Interdisciplinarity". In: *Nature News* 525.7569, p. 313.

Sanz-Menéndez, Luis, María Bordons, and M. Angeles Zulueta (2001). "Interdisciplinarity as a Multidimensional Concept: Its Measure in Three Different Research Areas". In: *Research Evaluation* 10.1, pp. 47–58.

Schreibman, Susan, Laura Mandell, and Stephen Olsen (2011). "Introduction". In: *Profession* 2011.1, pp. 123–201.

"Facilitating Interdisciplinary Research" (2005). In: ed. by National Academy of Sciences and National Academy of Engineering. OCLC: 698599131. URL: http://bit.ly/FacilitatingIDR.

Shriver, Gene (2011). *Linking Language to the Technology and Communication Process*. URL: https://www.gala-global.org/publications/linking-language-technology-and-communication-process-0 (visited on 07/23/2018).

Simmons, Vani N. et al. (2011). "Transcreation of Validated Smoking Relapse-Prevention Booklets for Use with Hispanic Populations". In: *Journal of health care for the poor and underserved* 22.3, p. 886.

Sinclair, John (2005). "Corpus and Text-Basic Principles". In: *Developing Linguistic Corpora: A Guide to Good Practice*. Ed. by Martin Wynne, pp. 1–16.

Smith, Karen (2006). "Rhetorical Figures and the Translation of Advertising Headlines". In: *Language and Literature* 15.2, pp. 159–182. URL: http://journals.sagepub.com/doi/abs/10.1177/0963947006063745.

Smith, Karin Sim (2017). "On Integrating Discourse in Machine Translation". In: *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 110–121.

Solomon, Felicia M. et al. (2005). "Development of a Linguistically and Culturally Appropriate Booklet for Latino Cancer Survivors: Lessons Learned". In: *Health promotion practice* 6.4, pp. 405–413.

Somers, Harold (2001). "EBMT Seen as Case-Based Reasoning". In: *MT Summit VIII Workshop on Example-Based Machine Translation, Santiago de Compostela, Spain*, pp. 56–65.

— (2003). *Computers and Translation: A Translator's Guide*. John Benjamins Publishing. 369 pp. ISBN: 978-90-272-9669-6.

Spencer, Larry (2018). "Getting oldWeather Data Ship-Shape for Science". In: *Old Weather Blog*. URL: https://blog.oldweather.org/2018/12/04/getting-oldweather-data-ship-shape-for-science/ (visited on 12/29/2018).

Steiner, Christina M. et al. (2014). "Evaluating a Digital Humanities Research Environment: The CULTURA Approach". In: *International journal on digital libraries* 15.1, pp. 53–70.

Stirling, Andy. "Disciplinary Dilemma: Working across Research Silos Is Harder than It Looks | Andy Stirling". In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/science/political-science/2014/jun/11/science-policy-research-silos-interdisciplinarity (visited on 02/11/2019).

Szostak, Rick. *Defining "Multidisciplinary" and "Cross-Disciplinary"*. URL: https://sites.google.com/a/ualberta.ca/rick-szostak/research/about-interdisciplinarity/definitions/defining-multidisciplinarity-and-cross-disciplinarity (visited on 12/15/2018).

— *Defining "Transdisciplinary"*. URL: https://sites.google.com/a/ualberta.ca/rick-szostak/research/about-interdisciplinarity/definitions/defining-transdisciplinarity-and-multidisciplinarity (visited on 12/15/2018).

— *History of Disciplines and Interdisciplinarity*. URL: https://sites.google.com/a/ualberta.ca/rick-szostak/research/about-interdisciplinarity/history-of-interdisciplinarity (visited on 11/12/2018).

— (2002). "How to Do Interdisciplinarity: Integrating the Debate". In:

Szostak, Rick and Pauline Gagnon (2013). "The State of the Field: Interdisciplinary Research". In: *Issues in interdisciplinary studies*.

Sá, Creso M. (2008). "'Interdisciplinary Strategies' in US Research Universities". In: *Higher Education* 55.5, pp. 537–552.

TEI. "TEI: Guidelines". In: URL: http://www.tei-c.org/Guidelines/ (visited on 11/14/2016).

The Eighth Amendment of the Constitution - The Citizens' Assembly. URL: https://www.citizensassembly.ie/en/The-Eighth-Amendment-of-the-Constitution/ (visited on 01/29/2019).

Thompson, Jessica Leigh (2009). "Building Collective Communication Competence in Interdisciplinary Research Teams". In: *Journal of Applied Communication Research* 37.3, pp. 278–297.

Torres-Simón, Esther and Anthony Pym (2016). "The Professional Backgrounds of Translation Scholars. Report on a Survey". In: *Target. International Journal of Translation Studies* 28.1, pp. 110–131.

Tress, Bärbel et al. (2003). *Interdisciplinary and Transdisciplinary Landscape Studies: Potential and Limitations*. Delta Program Wageningen.

Turcato, Davide and Fred Popowich (2003). "What Is Example-Based Machine Translation?" In: *Recent Advances in Example-Based Machine Translation*. Springer, pp. 59–81.

Van Dijk, Teun A. (1985). "Handbook of Discourse Analysis". In: *Discourse and Dialogue*. Citeseer.

Van Noorden, Richard (2015). "Interdisciplinary Research by the Numbers". In: *Nature News* 525.7569, p. 306.

Viseu, Ana (2015). "Integration of Social Science into Research Is Crucial". In: *Nature* 525.7569, pp. 291–291. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/525291a. URL: http://www.nature.com/doifinder/10.1038/525291a (visited on 08/01/2016).

Wallace, Danny P. (2007). *Knowledge Management: Historical and Cross-Disciplinary Themes*. Libraries unlimited.

Wang, Longyue et al. (2016). "A Novel Approach to Dropped Pronoun Translation". In: *arXiv preprint arXiv:1604.06285*.

Way, Andy (2018). "Quality Expectations of Machine Translation". In: *arXiv preprint arXiv:1803.08409*.

Weather, Old. *Old Weather - About*. URL: https://www.oldweather.org/about.html (visited on 12/17/2018).

Weingart, Peter (2010). "A Short History of Knowledge Formations". In: *The Oxford handbook of interdisciplinarity*, pp. 3–14.

Wells, Kristen J. et al. (2013). "Feasibility Trial of a Spanish-Language Multimedia Educational Intervention". In: *Clinical Trials* 10.5, pp. 767–774.

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *arXiv preprint arXiv:1609.08144*.

Wynne, Martin (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Vol. 92. Oxbow Books Oxford.