Syntactic Variation of Complementisers Among Irish Dialects

Author: Daniel Gallagher
Date: October 13, 2024

1 Introduction

This work aims to provide a comprehensive empirical analysis of Irish complementiser patterns among its three primary dialects: Connacht, Munster, and Ulster. The modelling of these patterns is based on the work of Mccloskey 2002 and classifies each pattern as having (1) no long-distance dependencies, (2) successive cyclic movement, or (3) a resumptive pronoun. Additionally, an effort is made in this work to bridge theoretical linguistics and empirical research by building tools for the algorithmic analysis of these patterns on a large dataset. The various types of complementiser patterns are introduced in Section 2 and a number of syntax trees are built with a particular focus on demonstrating an interpretation of successive cyclic movement in Irish. Data is collected from the National Corpus For Ireland and pre-processed to include only sentences which contain a tag for their dialect and can be identified as having a complementiser in them, meaning there is at least one matrix clause and one embedded clause in every sentence. This amounted to a final dataset size of 341,193 sentences. Given that Irish is a minority language and relatively understudied, computational tools for natural language processing (NLP) are not as readily available as for major languages such as English, German, or Spanish (Lohar et al. 2023). Therefore, a tool for cyclically parsing Irish sentences into their clauses and complementiser structure was built by the author for this work using *Python*, a high-level programming language, and *SpaCy*, open-source tools for NLP (Honnibal et al. 2020). The parser was designed so as to model a similar process of cyclicity in theoretical linguistics, acting on only one level of embedding at a time and recursively iterating over the entire sentence structure. Each complementiser is then parsed and tagged with one of McCloskey's tags for the four types of complementiser patterns. To the author's knowledge, it is the first such cyclic syntactic parser and complementiser tagger for Irish. This parser was made open-source and is publicly available at *github.com/DanielGall500/irish-syntax-parser*. The results of running this on the dataset are provided in Section 6 and an analysis is provided in Section 7 which aims to provide insight into how complementiser pattern usage varies across the dialects of Irish and how the language is changing in this respect as it evolves.

2 Complementiser Forms in Irish

Irish has three primary categories of complementiser - *go*, *aN*, and *aL* (Mccloskey 2002). What is particularly interesting about these is how they use differing marking strategies to reveal a clear distinction between a lack of movement and successive cyclic non-local $\bar{A}$ movements. The structure of these finite complementiser clauses is sensitive to $\bar{A}$-extractions from the clause of which they are the head. The first complementiser is used in cases where there is no long-distance dependencies, while the other two (aL, aN) employ two different strategies to model long-distance dependencies.

2.1 No Long-Distance Dependency Pattern

We begin with the *go* particle for finite complement clauses, which is the simplest case where there are no long-distance dependencies (absence of A-binding). This form will be therefore referred to as No-LD. The *go* particle will typically be preceded by a verb ('He said that...') or adjective ('I am sure that...'). This particle can surface in distinct forms to agree with tense, for instance in sentence 1 the particle *go* takes a past-tense marking to become *gur*.

(1) Creidim gu-r inis sé bréag.
believe.1SG *go*.PST tell.PST she lie
'I believe that she told a lie.' (Mccloskey 2002: p. 2)

The forms of this particle are 'go', 'gur' (past / conditional form when followed by a consonant), 'gurb' (present / future form before a vowel) , and 'gurbh' (past / conditional form when followed by a vowel). Note that the use of this particle triggers lenition.

2.2 Gap Pattern

Any finite clause which contains a trace due to movement to an $\bar{A}$ position is introduced using a different particle, one which is termed *aL*. Unlike the previous particle, this always surfaces as 'a' (see sentences 2 and 3). This form of complementiser structure will be referred to as GAP. This form is a clear example of successive cyclic movement occurring across phases.

(2) an t-ainm a hinnseadh dúinn a bhí __ ar an áit.
the name *aL* tell.PST.PASS us *aL* is.PST t on the place
'The name that we were told was on the place.' (Mccloskey 2002: p. 3)

(3) An fhilíocht a chum sí __.
the poetry *aL* compose.PST she t
'The poetry that she composed.' (Mccloskey 2002: p. 4)

Where there is movement across multiple clauses this particle must be used to introduce each respective clause. This can be observed in sentence 4.

(4) cuid den fhilíocht a chualaís ag do sheanmháthair arán a cheap an sagart úd __
some of-the poetry *aL* heard by your grandmother being-said *aL* composed the priest DEMON t
'Some of the poetry that you heard your grandmother saying that that priest composed.' (Mccloskey 2002: p. 8)

2.3 Resumptive Pattern

We previously saw how the *aL* particle is used to model a long-distance dependency, the strategy being that there is a trace left from movement across a finite clause to an $\bar{A}$ position. The *aN* particle is employed in a similar manner, though the gap is instead filled by a resumptive pronoun. This form will be referred to as RES. The particle can be realised in two different surface forms - 'a' or 'ar' - the latter being used in the case of the selected clause being marked in the past tense. Sentence 5 shows an example of the resumptive pronoun 'her' referencing back to 'the girl'.

(5) An ghirseach a-r ghoid na síogaí í.
the girl *aN*.PST steal.PST the fairies her
'the girl that the fairies stole away.' (Mccloskey 2002: p. 7)

Unlike the previous particle *aL*, the particle *aN* does not appear for each clause across which movement took place. The least marked pattern will use the *aN* for the head C and the *go* particle is used for each finite clause which follows. See example sentences 6 and 7.

(6) An t-ór seo ar chreid corr-dhuine go raibh sé ann.
the DEF gold DEM PST believe.PST some-person COMP be.PST 3SG there
'The gold that some people thought (it) was there.' (Mccloskey 2002: p. 9)

(7) Mícheál sin a raibh dóchas acu uilig go bpósfadh sé Róise
Mícheál DEM AN be.PST.3SG hope at-3PL all GO marry.COND 3SG Róise
'that Mícheál that they all hoped (that he) would marry Róise.' (Mccloskey 2002: p. 15)

2.4 Modern Resumptive Pattern

A resumptive form which has appeared in more recent Irish usage is that of the modern resumptive, or MODERN-RES as it will be categorised in this work. This form uses the *go* particle in place of *aN* for the head C in resumptive structures. It is briefly claimed to occur in Munster varieties and some Southern Connacht varieties in Mccloskey 2002. They do not give this form a separate particle tag, however we will introduce here the *goN* tag for this more modern resumptive pattern in order to distinguish it from the standard *go* particle.

(8) Tigh beag caol gu-r mhaireamar ann.
house little narrow go-PST live.PST.3SG in-it
'It was a narrow little house that we lived in.' (Mccloskey 2002: p. 30)

(9) an simné gu-r dhóigh leat go leagfadh gálaí an gheimhridh é
the chimney go-PST think.PST with-you go would-knock-over gales the winter.GEN it
'the chimney that you'd think the winter gales would topple' (Mccloskey 2002: p. 30)

2.5 Syntax of Complementiser Patterns

With this introduction in Section 1 of the four primary complementiser patterns in modern Irish introduced originally in Mccloskey 2002, we can turn towards a brief investigation of what sort of syntactic properties we may observe here and build trees to demonstrate these properties.

The GAP structure is particularly interesting as we must explain how this long-distance movement can take place. There are two alternative lines of reasoning for deriving how this particle is realised. In each case there is a relation between the *movement* of a DP to a non-local $\bar{A}$ position and the *complementiser head* (C-head). The first line of reasoning is:

1. Movement takes place of a DP to a non-local $\bar{A}$ position.
2. Due to this movement a mark is deposited on a the head C.

The second line of reasoning is somewhat a reversing of the logic:

1. The head C contains a property which forces movement to take place to a non-local $\bar{A}$ position.
2. Movement takes place due to this property.

We will assume that the latter is true in this case, with SPEC-CP containing a probe that is attracting a DP and leaving a trace in its place. We do not assume the existence of little v following McCloskey 2022, nor do we assume that TP is a phase. For the construction of the proper derivations, there are three important distinctions between the various types of complementiser patterns that must be made. In the simplest case, we have the NO-LD pattern where there are no long-distance dependencies. Here we do not see MOVE or MERGE occur. In the GAP pattern we see $\bar{A}$ movement taking place and a trace being left in its place. We will assume that this trace is left in SPEC-CP. Lastly, the RES form occurs as a result of the merging of the resumptive pronoun as a repair strategy. Note that the MODERN-RES pattern will not be accounted for her as it holds the same properties as RES. Keeping these important points in mind, in examples 10, 11, and 12 we see three data points which were written by the author and glossed with the Leipzig glossing rules (MPI 2015). These will be used to illustrate these patterns.

(10) Creidim go feicim an madra.
believe.PRES.1SG GO see.PRES.1SG DEF dog
'I believe that I see the dog.'

(11) Feicim an madra a dúirt Jack a chonaic sé.
see.PRES.1SG DEF dog AL say.PST.3SG Jack AL see.PST.3SG he
'I see the dog that Jack says that he saw.'

(12) Feicim an madra a dúirt Jack go chonaic sé é.
see.PRES.1SG DEF dog AN say.PST.3SG Jack GO see.PST.3SG he he.ACC
'I see the dog that Jack says that he saw (it).'

Figure 1 shows the syntactic tree for sentence 10, figure 2 for sentence 11, and figure 3 for sentence 12. These depict one potential interpretation of the syntactic structure of the NO-LD, GAP, and RES patterns, respectively. Irish has a verb-initial word order (VSO) and as such the finite verb comes in the initial position. For instance, "I am" in Irish is "tá mé", literally "am I". Verbs can occur in two forms, one which contains the verb plus a pronoun occurring after the verb as in the previous example, and another in which is inflected for tense and subject as in *creidim* "I believe". In order to model this word order we assume that the verb originates in VP and head movement occurs from VP to

TP in order to check the strong features of T, that is, tense and agreement features. We can additionally assume that all subjects originate in VP and move to their surface position following the VP-internal subject hypothesis. Figure 2 represents successive cyclic movement of the DP *an madra* "the dog" through multiple clauses. It moves from its original position as the object in the most embedded clause through Spec-CP twice arrives at its final landing site Spec-VP in the matrix clause. These movements leave behind the traces t_i in two locations. This successive cyclic movement is marked by the *aL* complementiser in Irish. In figure 3 the Res pattern is observed where successive cyclic movement is avoied through the merging of the resumptive pronoun *é* "he" which is referring to the DP *an madra* "the dog". This avoids the high level of complexity involved in forming the Gap pattern.

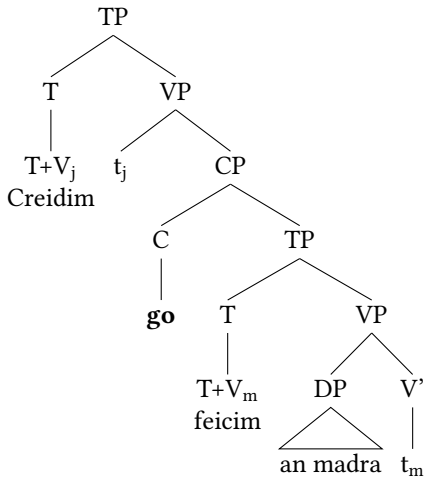


Figure 1: Syntactic Modelling of No-LD Pattern

3 Research Questions

There are hence a number of patterns of complementiser usage which we see in Irish that each display varying morphological realisations. These patterns are not as stringent as they once were however. The vast majority of modern Irish speakers have learned Irish as a second language and it is being constantly morphed and altered through each generation. Additionally, there are strong dialectal variations in Irish which contribute to varying results when speakers are asked to produce sentences in which we would expect to see certain patterns as they are listed above. McCloskey notes that Munster varieties and some Southern Connacht varieties have begun to use the complementiser *goN* in place of the complementiser *aN* in resumptive structures, for instance. He claims that some speakers, particularly of the younger generation, are casting aside the distinction between *aN* and *aL* as well as using the Modern-Res pattern in place of Res. The patterns which will concern us most are outlined in Table 1. As clarified in Section 2, these patterns will be referred to throughout as No-LD, Gap, Res, and Modern-Res, respectively. There are ways in which these patterns can combine and appear in different forms which we will not discuss here such as in mixed chain types.

Description	**Pattern**
No long-distance dependencies	VP $[_{CP}$ *go* . . .]
Gap structure	XP$_j$ $[_{CP}$ *aL* . . . $[_{CP}$ *aL* . . . $[_{CP}$ *aL* . . . t_j . . .]]]]
Resumptive structure	DP $[_{CP}$ *aN* . . . $[_{CP}$ *go* . . . *pro* . . .]]
Modern resumptive structure	DP $[_{CP}$ *go* . . . $[_{CP}$ *go* . . . *pro* . . .]]

Table 1: Irish Complementiser Patterns (Mccloskey 2002)

We have observed that Irish has three distinct complementisers which each have unique realisations in the language. We have also touched upon how for one of these complementisers - *aN* - McCloskey claims that there

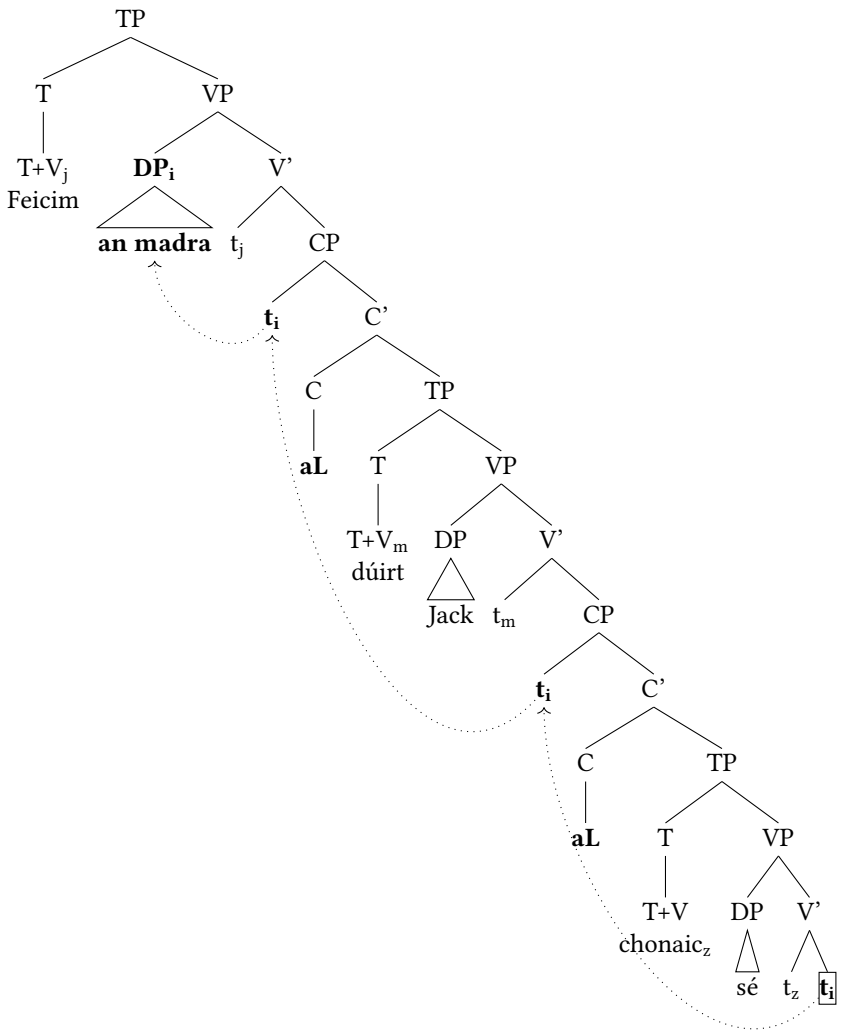


Figure 2: Syntactic Modelling of Gap Pattern

is a notable fall in usage, instead resorting to the use of *go* in its place. This leads us to the question of how these complementisers may be changing in their usage in modern Irish. More formally put, we can ask:

1. What is the relative usage frequency for the various complementiser patterns in modern Irish?
2. Is there a tendency among certain dialects to lose the distinction between aL (non-resumptive) and aN (resumptive)?
3. Do certain dialects of modern Irish tend to use *go* in place of *aN* in resumptive structures?

4 Data

This project aims to provide a comprehensive quantitative analysis of complementiser usage patterns among various dialects of Irish. Therefore, it was imperative to access high-quality data which was categorised by dialect. The data used for this work was collected from the *New Corpus For Ireland* (NCI) (Kilgarriff, Rundell & Dhonnchadha 2006), where a large collection of diverse Irish texts collected and stored with the intention for linguists to carry out research. The NCI originally began as a database of lexical information for creating English-Irish dictionaries, but

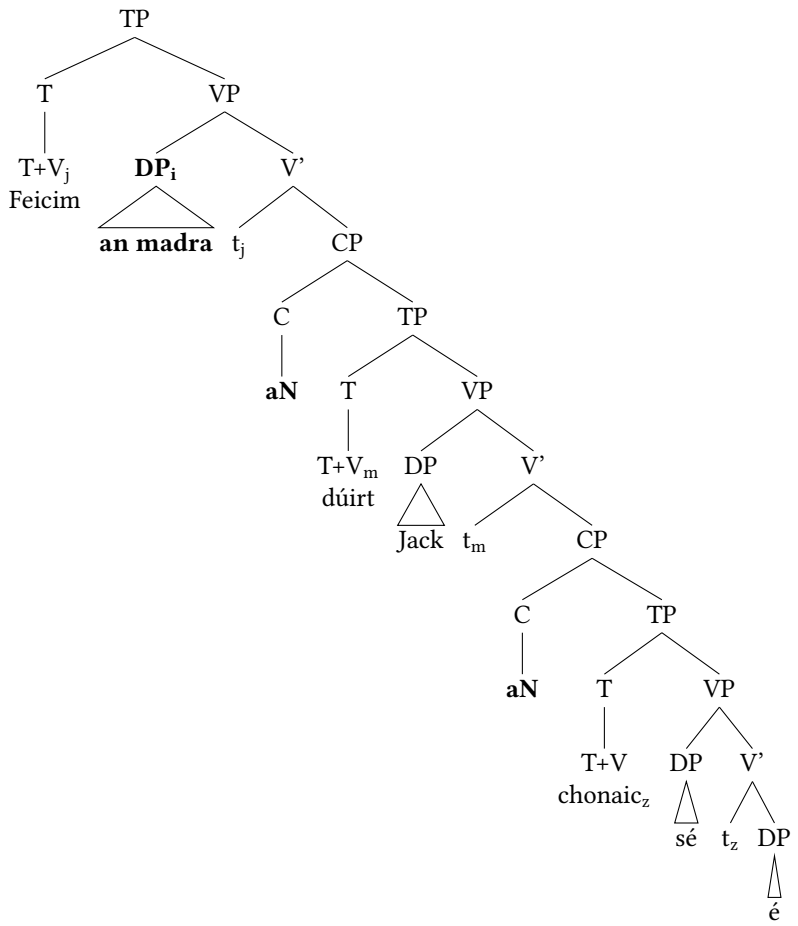


Figure 3: Syntactic Modelling of Resumptive Pattern

has since expanded to include many more types of Irish texts and with a very powerful interface for data filtering. The database cannot be publicly accessed but this was achieved thanks to correspondence with the creators.

4.1 Dataset Creation

Six datasets were created in total based on the three primary Irish dialects of Connacht, Munster, and Ulster as well as on whether the sentence included potentially included (1) the *go* particle or (2) the *aL* or *aN* particle. This initial dataset included false positives, but these will be recognised as such and discarded during the processing phase. The resulting datasets created from the NCI corpus for the analysis as well as POS-tagging can be viewed in Table 2 and 3, respectively. The size of the datasets before processing and after processing can be viewed in the pre- and post-feature columns.

4.1.1 Querying for Connacht, Munster, and Ulster Dialect

This corpus uses a query language in order to filter the dataset and retrieve the data that is relevant for one's particular research questions. In this case, six queries were required in order to construct the datasets. These accounted for dialect as well as complementiser surface variations. **Queries for Dataset Retrieval:**

```
Query:[word="a"]|[word="ar"] within <doc (dialect="Connacht")/>

Query:[word="a"]|[word="ar"] within <doc (dialect="Munster")/>

```

```
Query:[word="a"]|[word="ar"]
    within <doc (dialect="Ulster")/>

Query:[word="go"]|[word="gur"]|[word="gurb"]|[word="gurbh"]
    within <doc (dialect="Connacht")/>

Query:[word="go"]|[word="gur"]|[word="gurb"]|[word="gurbh"]
    within <doc (dialect="Munster")/>

Query:[word="go"]|[word="gur"]|[word="gurb"]|[word="gurbh"]
    within <doc (dialect="Ulster")/>
```

For each query a maximum of 100,000 sentences were collected. A number of pre-processing steps were implemented in order to improve the quality of the data, for instance removing duplicates or irrelevant words such as indefinite or definite articles. The operations performed are dependent on your research question, and so in this case we perform a number of operations which we deem relevant for analysing complementiser constructions in Irish. The operations implemented in this case were:

- Removing duplicates from the dataset as this was likely an issue with data collection.
- Strings were reduced to a single sentence instead of spanning potentially multiple.
- Punctuation was removed.
- Any sentences which were discovered to not contain at least one complementiser were removed (e.g "go" can be a complementiser or intensifier, and so not all filtered data may be relevant.

These final processed datasets are summarised in Table 2. The final total number of sentences that will be analysed in this work is 341,193.

Dataset	Complementiser	Dialect	Pre Sentence Count	Post Sentence Count
aL_aN_Connacht_100k.csv	aL or aN	Connacht	100,000	51,314
aL_aN_Munster_100k.csv	aL or aN	Munster	100,000	52,657
aL_aN_Ulster_100k.csv	aL or aN	Ulster	100,000	48,452
go_Connacht_100k.csv	go	Connacht	80,000	63,446
go_Munster_100k.csv	go	Munster	80,000	56,538
go_Ulster_100k.csv	go	Ulster	80,000	68,786

Table 2: Datasets Created from Foclóir Corpus.

4.1.2 Word Category Identification

In order to develop the tools required to syntactically analyse a larger dataset of sentences, a basic tagger for word category was built, otherwise known as a *part-of-speech tagger*. The most relevant word categories for complementiser analysis were *nouns* and *adjectives*. These help determine whether an NP is modifying the complementiser clause and whether a complementiser such as "go" is truly a complementiser or really an intensifier, respectively. The collected datasets are summarised in Table 3.

Dataset	Word Category	Word Count
focloir_nouns.csv	Nouns	48316
focloir_adjectives.csv	Adjectives	6289

Table 3: Word Category Datasets for POS-tagging.

5 Methodology

A greater emphasis is being continuously put on improving *quantitative* analyses of syntactic theories rather than purely *qualitative* analyses, for instance computational experiments run to determine the validity of various analyses of the same syntactic phenomenon (Ermolaeva 2021). This work follows a similar philosophy and aims to focus on quantitative means in order to complement previous qualitative work on Irish complementiser patterns. To this end, a tool has been constructed which can carry out analyses of large amounts of Irish sentences and parse them for their complementiser constructions, thus allowing us to provide strong empirical evidence for any variations we may find.

5.1 Cyclic Irish Syntax Parser

An Irish complementiser parser was built for this work with the purpose of parsing Irish sentences into their individual clauses and mining them for syntactic phenomena. Taking inspiration from cyclicity in syntax, it iterates cyclically (or recursively) over the hierarchy of a given Irish sentence and stores the clauses as nested objects in a data structure. This falls in line with the theory of islandhood where movements that appear to take place non-locally are in fact a result of a series of local operations. This is illustrated in example 13. Along with this, the tool performs basic part-of-speech validity tests to check what sort of categories of word occur in various important positions in each clause such as clause-initially or -finally.

(13) $\text{XP}_j\ [_{Iteration1}\ t_j\ \text{C} \ldots [_{Iteration2} \ldots [_{Iteration3}\ t_j\ \text{C}\ [_{Iteration4}\ t_j \ldots]]]]$

5.1.1 Functionality

The parser performs the following tasks on a sentence S:

1. Various pre-processing steps such as lowercase conversion and the reduction of a group of semantically related items into a single form, known as lemmatisation.
2. Iterate through the clauses of S recursively by performing a complementiser identification process which analyses the structure of S, its lemmas and its part-of-speech tags. For instance, a category of type NUM (number) should not follow "go" as this may indicate that it is in fact an intensifier and not a complementiser.
3. Store the clauses of S and part-of-speech tags in a nested JSON structure. This includes information about the occurrence of resumptive pronouns.

5.1.2 Code

This tool was built using Python 3.9 and incorporated Irish lemmatisation functionality from the SpaCy lemmatiser (Honnibal et al. 2020). Given that others may like to carry out similar such research on Irish clause structure in the future, it is available for download on Github at *github.com/DanielGall500/irish-syntax-parser*. A small test dataset of 16 Irish sentences containing these complementisers has been compiled for initial testing purposes based off of the examples provided in Mccloskey 2002. Unit tests were set up to test whether the tool correctly identified complementiser clauses, word categories, as well as the occurrence of resumptive pronouns. Continuous Integration testing was then set up to run these tests each time any changes were made to the code.

5.1.3 Storing Clause Structure

The parser takes the original Irish sentence as its input and returns a JSON structure, which is a standard format in which data can be stored so that it can be understood well by both computers and humans. Some additional information also encoded in this data structure has been left out for clarity. Note that the parsed sentence has been lemmatised, so for instance *fhilíocht* becomes *filíocht* in order to remove Irish lenition for instance. This is necessary here as it makes it easier to recognise a word's category which aids in parsing the syntactic structure of the sentence algorithmically. One such JSON structure is shown in the following example. We can see that the structure comprises of *headers*, which indicate the title of a particular feature, and *values*, which store the data that we are interested in.

Irish Sentence: "*An fhilíocht a chum sí.*"

English Translation: *"The poetry that she composed."*

Parsed Sentence: "[An filíocht [a cum sí]$_{no\_resumptives}$]"

JSON Data Structure:


```
{
    "full": "An fhilíocht a chum sí.",
    "lemmas": [
        "an",
        "filíocht",
        "a",
        "cum",
        "sí"
    ],
    "clause_structure": {
        "clause": [
            "an",
            "filíocht"
        ],
        "selected_complementiser": "a",
        "embedded_clause": {
            "clause": [
                "cum",
                "sí"
            ]
        }
    },
    "num_clauses": 2
},
```


5.2 McCloskey Parsing

For this research it was also desirable to be able to convert these representations into a format such as that which was provided in Table 1. For instance, a sentence such as "Creidim gur inis sé bréag", meaning "I believe that he told a lie" would be parsed into *XP go []*, indicating that it contains an XP with a single *go* complementiser clause. It was particularly important that the distinction between *go*, *aL*, *aN*, and *goN* be observed in the resulting patterns as was described in Mccloskey 2002. This functionality was implemented and termed *McCloskey parsing*. A full example can be seen in Table 4, where the 20 most common patterns in the aL/aN Connacht dataset are shown. We can see for instance that the GAP pattern was the most popular in this dataset. Identification of these structures is no simple task and this functionality should not be taken as fully accurate, however future work can aim to improve these tools and provided a more comprehensive insight into their usefulness for linguistic research. For the current version of the McCloskey parser, only a distinction between XP and DP was observed as these were most important.

6 Results

The matter at hand is now to run these computational tools for syntactic analysis on the sizeable Irish dataset, which incorporates a total of 341,193 sentences, and produce a number of visualisations and representations of the data which will aid our analysis in Section 7 and inform the answers to the research questions laid out in Section 3. Each result is typically divided into individual dialects and where applicable results include a maximum of 3 embedded clauses to limit the sentence complexity which was incorporated into the final results.

Pattern	Sentence Count
XP aL [ ]	13933
DP aL [ ]	10587
XP aL [aL [ ]]	4430
DP aL [aL [ ]]	3598
XP aN [ ]	3380
DP aN [ ]	3046
XP go [a [ ]]	1604
XP aL [aL [aL [ ]]]	1132
DP go [a [ ]]	1113
DP aL [aL [aL [ ]]]	991
XP go [ ]	833
XP a [go [ ]]	655
DP go [ ]	532
DP a [go [ ]]	450
XP go [a [a [ ]]]	436
XP aL [aL [aL [aL [ ]]]]	346
DP go [a [a [ ]]]	328
DP aL [aL [aL [aL [ ]]]]	311
XP aN [go [ ]]	297

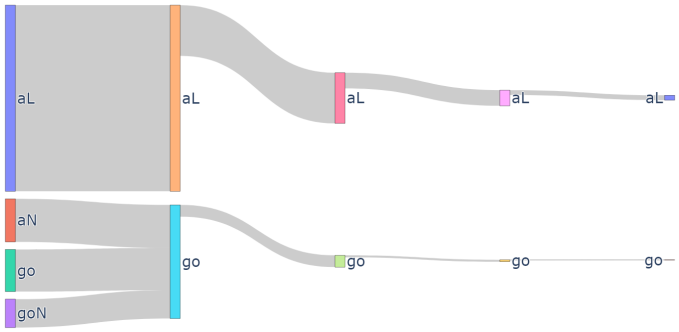
Table 4: Top 20 results from running McCloskey parsing on aL/aN Connacht Dataset.

6.1 Tables

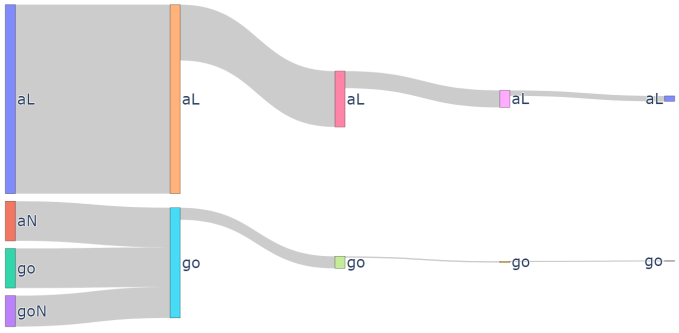
The first major result is the distribution of the various complementiser patterns among dialects as outlined in Table 5. This is intended to model Table 2, created in Mccloskey 2002, but with the percentage occurrence of various patterns by dialect. In order to gain a closer look at the most frequent patterns for each dialect overall, Table 6 outlines the top 5 constructions from each. We observe that the most common constructions are typically a complementiser with a single embedded clause. The No-LD and Gap patterns are the most frequently used in each case, regardless of dialect. The occurrence of a resumptive pronoun or lack thereof (indicating a trace) is important for research question 2 and hence an overview of the frequency of resumptives in the final clause of a sentence is given in Table 7.

6.2 Visualisations

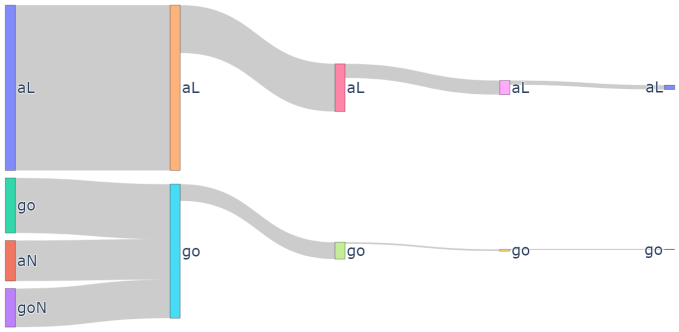
A Sankey diagram is a type of visualisation which typically represents the flow of a limited supply of resources or materials through a system with individual stages. However, this type of diagram has an application in many different fields. Figure 4 shows three Sankey diagrams which represent the flow of complementisers from one clause to another. Each stage in the diagram (from left to right) represents a new complementiser clause and the complementiser is indicated in text beside the stage barrier. The stream will always get smaller as one moves from left to right, as the initial stage includes *any sentence which contains at least one complementiser*, which in the case of this dataset is all of them. The next stage is *any sentence which contains at least two complementisers*, and then the next stage *...at least three complementisers*, and so on. It is immediately apparent that the three dialects have a very similar flow of complementisers for the data collected, and therefore the final results are not going to provide huge variations in usage but rather more subtle differences.



(a) Connacht Data



(b) Munster Data



(c) Ulster Data

Figure 4: Clause-by-Clause Flow of Complementisers

Category	Pattern	Connacht	Munster	Ulster
No-LD	VP $[_{CP}$ *go* . . .]	27.05%	25.80%	26.61%
Gap	XP_j $[_{CP}$ *aL* . . . $[_{CP}$ *aL* . . . $[_{CP}$ *aL* . . . t_j . . .]]]]	30.21%	31.57%	27.99%
Res	DP $[_{CP}$ *aN* . . . $[_{CP}$ *go* . . . *pro* . . .]]	4.29%	4.03%	3.95%
Modern-Res	DP $[_{CP}$ *go* . . . $[_{CP}$ *go* . . . *pro* . . .]]	0.35%	0.39%	0.58%

Table 5: Distribution of linguistic features across Irish regions (max 3 embedded clauses)

Pattern	Sentences	%
XP go [ ]	16607	14.47%
XP aL [ ]	13933	12.14%
DP go [ ]	13608	11.86%
DP aL [ ]]	10587	9.23%
XP go [aL [ ]]	5046	4.4%

Connacht

Pattern	Sentences	%
XP go [ ]	15440	14.13%
XP aL [ ]	14249	13.05%
DP go [ ]	11821	10.83%
DP aL [ ]	9594	8.79%
XP aL [aL [ ]]	4727	4.33%

Munster

Pattern	Sentences	%
XP go [ ]	16751	14.29%
DP go [ ]	13710	11.69%
XP aL [ ]	13207	11.27%
DP aL [ ]	9651	8.23%
XP go [aL [ ]]	5402	4.61%

Ulster

Table 6: Frequency of Complementiser Patterns By Dialect

Dataset	Region	Sentences	Final Clause Resumptives	Percentage
aN/aL	Connacht	51314	10226	19.93%
aN/aL	Munster	52657	10625	20.18%
aN/aL	Ulster	48452	9380	19.36%

Table 7: Usage of Resumptives in aN/aL Dataset

7 Analysis

We began with a dive into the theoretical space for Irish complementiser patterns, however the results in Section 6 provide us with the necessary information in order to understand better the real-world usage space. We can now turn to the original research questions of this work outlined in Section 3 and attempt to provide a number of insights from these results to shed light on each.

7.1 Variation of Complementiser Patterns Among Dialects

The first question we will concern ourselves with us that of the usage of various complementiser patterns and their variation among different Irish dialects. Table 6 shows us that the four most frequent patterns are consistent across all dialects. That is, "no long-distance dependency" (NO-LD) structures and "gap" structures occur most often within the dataset, regardless of dialect. Interestingly, we observe that a DP is not particularly less likely to occur directly before a No-LD structure than for a Gap structure. In the case of Connacht and Ulster we see the construction *XP go [aL [ ]]* is the fifth most popular at 4.4% and 4.61%, respectively. This is a construction such as "I thought that I ordered the lamp that he recommended". In the case of Munster however the 5th most-common is *XP aL [aL []]*, for instance "I saw the electrician that recommended the lamp that I ordered." Turning towards an analysis of the four patterns as a whole, Table 5 gives an overview of this important distribution within our dataset. While the data does show relatively similar usage patterns of the patterns, we do observe a number of variations. If we look at the usage of aL/aN complementisers, the Ulster dialect tends to use less Gap, and Resumptive patterns than its counterparts. There are two ways to account for this, (1) the Modern-Res pattern accounts for a higher percentage than in other dialects of usage and (2) the data taken from the Ulster region may have contained longer sentences with more levels of embedding which was not accounted for in the data, as the maximum of 3 embedded clauses accounted for in the final distribution does not show the entire picture. The Munster dialect is more likely than others to use a Gap pattern, particularly in comparison with that of Ulster. It is clear that Connacht is more likely to use a Gap or Res structure and less likely to use a Modern-Res, while Ulster stands in contrast to that with a higher amount of Modern-Res and less Gap and Res patterns. The usage of Modern-Res will be discussed further in Section 7.3. What can additionally be noted is that in no dialect is Modern-Res close to overtaking the standard Res. One reason for this however may be that the data was not collected from informal contexts such as texts and social media posts. In these contexts it may be more frequent to use non-standard constructions such as these and further work may look at this in the future.

7.2 Loss of Resumptive/Non-Resumptive Distinction

This research question focuses on the Gap and Res constructions specifically and how their usage may be merging over time. In order to examine this distinction we will look specifically at the aL/aN datasets which were created with the intention of being able to focus on these long-distance dependency patterns. Table 7 show us the number of instances among Gap (aL), Res (aN) and Modern-Res (goN) constructions where a resumptive pronoun appears in the final clause. We can see that the results are quite consistent across dialects with approximately one fifth of each dialect containing a resumptive pronoun. In Table 5 we observe that the Gap structure is typically far more common than its resumptive and modern resumptive counterparts. Therefore, when a speaker is forming a long-distance dependency, it is clear that a distinction between resumptive and non-resumptive is still very much observed in the data. The Gap structure does however remain by far the most common of the long-distance dependency patterns and it may well be that resumptive patterns are on a downward trend.

7.3 Usage of the Modern Resumptive Pattern

An important matter at the heart of this research is the variation in usage of the modern resumptive pattern defined in Table 2. This pattern uses the same structure as a typical Res structure, however the head C is filled by *go* instead of aN. Mccloskey 2002 claims that there is evidence for pattern variation among dialects for this pattern in particular, with Munster varieties and some southern Connacht varieties using this more recent form of resumption structure. However, the data collected for this research provides some evidence to the contrary. We see that in Table 5 that the Munster and Connacht dialects are less likely to use this form than the Ulster dialect and more likely to use a traditional Res pattern. The distribution is also consistent; there is a 0.23% difference in the likelihood of using Modern-Res between Connacht, the least likely, and Ulster, the most likely, which is somewhat balanced out by the 0.34% difference in usage of the more typical Res. This means that in the cases where a dialect will use a resumptive pronoun to model a long-distance dependency (Res or Modern-Res), the Ulster dialect will use the modern *goN* particle for the head C 12.8% of the time, while Munster 8.82% and Connacht 7.54%. This stands in direct contradiction to McCloskey's claim and we would therefore suggest that further work is undertaken to examine the usage patterns among dialects and how exactly they are distributed.

8 Conclusion

Modern Irish is an evolving language due to its recent boost in popularity through online platforms and communities forming such as the "Pop up Gaeltacht" (one-night events in pubs which encourage the speaking of Irish in conversation and with staff). The implications of this are that the language changes year-on-year and speakers are continuously adopting new expressions and constructions. For this reason, even our most basic assumptions about the syntactic constructions that modern-day Irish speakers are using must be examined. This work provided one such investigation into the usage of complementiser patterns across Irish dialects. This work built on top of work from Mccloskey 2002 and aimed to cross-compare dialects in this respect as well as provide insight into how the language might be changing with regard to resumption usage and which complementiser is used for the head in resumptive patterns. Access was granted to the NCI Irish corpus and using a number of queries to isolate complementiser candidates outlined in Section 4.1.1 we collected a total of 540,000 sentences. This initial dataset was pruned once again to include only sentences in which a complementiser was detected to a final size of 341,193 sentences. The Irish dialects of Connacht, Munster, and Ulster all came pre-tagged and hence each made up a particular proportion. We then created a number of open-source and publicly available natural language processing tools for syntactic analysis of Irish. The goal of this was in particular to be able to parse a large number of sentences into a form which would explicitly outline not only complementiser structure but the details of whether a *go*, *aN*, *aL*, or *goN* complementiser is being used for any given clause. This was a particularly difficult challenge due to a number of linguistic subtleties such as whether a pronoun is resumptive or not and additional problems of data pre-processing such as whether a data sample includes one whole sentence or multiple. For these reasons these tools should not be taken as fully polished but rather as a start point for quantitative analyses of syntactic phenomena in Irish. One potential avenue for future work here is the usage of the tagged data in order to improve part-of-speech taggers for Irish, particularly with respect to complementiser tagging. The final results of the analysis had a number of interesting insights. Sentences with a single embedded clause and a No-LD or Gap structure were by far the most common as outlined in Table 6. Res structures are most common in Connacht, followed by Munster and lastly Ulster. We observed that

when a speaker intends to use a resumptive structure, they will typically use GAP 80% of the time. Overall, the GAP pattern is the most common construction across dialects in order to model long-distance dependencies, and future work may look more closely at how the usage of resumptive structures has changed in recent years with younger speakers as it is relatively infrequent within the dataset. Contrary to McCloskey's claim, MODERN-RES was most common in the Ulster dialect amounting to 12.8& of the time in resumptive structures while notably less common in the Munster and Connacht dataset at 8.82% and 7.54%, respectively. This pattern is notably less common than all others standing at less than 1% of the total data for each dialect and so is far from overtaking other forms of modelling long-distance dependencies. However, care should be taken with assumptions here as younger speakers may not be well represented within the dataset.

References

Ermolaeva, Marina. 2021. Learning syntax via decomposition. UChicago. 169. https://doi.org/https://doi.org/10.6082/uchicago.3015.

Honnibal, Matthew et al. 2020. spaCy: Industrial-strength NLP. https://spacy.io/. Version 2.3.5.

Kilgarriff, Adam, Michael Rundell & Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the new corpus for ireland. Language Resources and Evaluation 40(2). 127–152. https://doi.org/10.1007/s10579-006-9011-7.

Lohar, Pintu et al. 2023. Building neural machine translation systems for multilingual participatory spaces. Analytics 2(2). 393–409. https://doi.org/10.3390/analytics2020022.

Mccloskey, James. 2002. Resumption, Successive Cyclicity, and the Locality of Operations. In chap. 8, 184–226. John Wiley Sons, Ltd. https://doi.org/https://doi.org/10.1002/9780470755662.ch9.

McCloskey, Jim. 2022. The syntax of irish gaelic. In Celtic languages and linguistics. Draft of a chapter for a volume (in preparation). Palgrave Macmillan.

MPI. 2015. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Accessed: 2024-10-13. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.