Unsupervised Labelling of Mutation Triggers in Welsh

Nicolás Gutiérrez-Rolón, Fernando Alva-Manchego
School of Computer Science and Informatics, Cardiff University, UK
{gutierrezrolonn, alvamanchegof}@cardiff.ac.uk

Abstract

Initial consonant mutation is a key feature of Welsh, but its complexity poses significant challenges for both language learners and natural language processing (NLP) systems. While existing tools can reliably detect mutated forms, they provide no information about why a mutation occurs, i.e. what grammatical or lexical factors trigger the change. This paper introduces the novel task of *mutation trigger labelling*, representing the first computational attempt to analyse and explain the reasons behind Welsh mutations. Two preliminary approaches are explored: (i) a linguistically-informed rule-based system integrating Constraint Grammar rules, and (ii) large language models (LLMs), prompted in few-shot settings. Our experiments test the feasibility of automatically identifying and labelling linguistic triggers behind Welsh mutations using a dataset constructed from grammar reference books and public corpora, and establish baseline insights into how context-aware mutation analysis can be achieved. By framing mutation trigger labelling as a linguistic computational problem, this work lays important groundwork within Welsh NLP and contributes to the broader development of explainable grammatical analysis for low-resource languages.

Keywords: Welsh mutation, mutation triggers, constraint grammar

1. Introduction

Welsh is a Celtic language belonging to the Brittonic subgroup, alongside Breton and Cornish. It is spoken by 19% of the population of Wales, approximately 580,000 people out of 3.1 million (Office for National Statistics, 2011). Welsh holds official minority language status under UK law, particularly the Welsh Language Act 1993 and the Welsh Language Measure 2011, which grants it official status alongside English in Wales. Welsh also has historical diasporic communities, notably in Patagonia, Argentina, where it has been maintained since the mid-19th century.

Despite ongoing revitalisation efforts, English continues to dominate as the de facto language in many domains (Davies, 2014). The Welsh Government works in promoting the language, ensuring it is an integral part of education, public services, and media, with the aim that "... persons in Wales should be able to live their lives through the medium of the Welsh language if they choose to do so" (Welsh Government, 2011).

The Welsh Government's Welsh language strategy, Cymraeg 2050: A million Welsh speakers (Welsh Government, 2017), has as one of its aims to "ensure that the Welsh language is at the heart of innovation in digital technology to enable the use of Welsh in all digital contexts". The Welsh Government's most recent Welsh Language Technology Action Plan further highlights the need for developing language technologies, and fostering a culture of open innovation (Welsh Government, 2018).

One of the most distinctive and challenging features of Welsh is initial consonant mutation, a morphophonological process where the first consonant of a word changes depending on its grammatical context. For example, a word like *teulu* (family) can look like *deulu*, *nheulu*, or *theulu* depending on its syntactic and lexical environment. While the rules governing this phenomenon are systematic, they are also highly context-dependent and influenced by complex interactions between syntax, morphology, and lexical factors.

Although fluent speakers generally apply mutation implicitly, explaining why a particular mutation occurs in a given sentence can be challenging even for proficient speakers. From a computational perspective, current Welsh NLP tools such as part-of-speech taggers can identify mutated forms with reasonable accuracy but cannot provide explicit causes or triggers of those mutations. This leaves a significant gap for both pedagogical and linguistic research purposes, as no existing computational system accounts for the different triggers that can cause mutation.

To our knowledge, this work represents the first attempt to define and operationalise the task of mutation trigger labelling in Welsh. Previous research has described mutation patterns linguistically (Ball and Müller, 1992; Mittendorf and Sadler, 2006; Awbery, 1986; Ball, 1984; Coupland and Ball, 1989), or used it to aid disambiguation (Tyers and Donnelly, 2009), but there has been no computational efforts to annotate, classify or explain their grammatical triggers. We therefore frame this paper as a pioneering first step toward building interpretable, context-aware mutation analysis resources. Such resources could enable learners to receive targeted grammatical feedback, facilitate automatic annotation in Welsh corpora, and enrich future research in syntax, morphology, corpus linguistics, and pedagogy.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11631–11641
11–16 May 2026.

By outlining the foundations of mutation trigger labelling and exploring potential computational approaches, this paper aims to open a new direction for Welsh NLP and for the study of morphosyntactic phenomena in low-resource languages broadly.

Our main contributions are:

- We define the novel task of mutation trigger labelling, establishing its linguistic motivation and computational significance.
- We present an initial framework for labelling mutation triggers, designing a taxonomy that represents underlying grammatical categories, functional contexts, stylistic constructions, and syntactic configurations.
- We explore two preliminary computational approaches to this task, a rule-based method using Constraint Grammar rules and an LLM prompting approach, as first attempts to model and evaluate mutation trigger labelling.[1]

2. Related Work

The Welsh language presents several challenges for NLP, due to its rich morphology and low-resource status. Even basic pre-processing tasks such as tokenisation and lemmatisation require dedicated tools, since resources trained on English fail to account for features such as different types of apostrophes and initial consonant mutations. As a result, Welsh NLP has had to rely heavily on rule-based linguistically-aware systems and handcrafted resources.

2.1. Early works

Early foundational work includes Apertium-cy (Tyers and Donnelly, 2009), a rule-based Welsh-English machine translation system. The authors highlighted the difficulties presented by initial-consonant mutations for tokenisation and tagging. Particularly, the authors noted that the HMM-based POS tagger was not able to take advantage of the disambiguation properties of mutation. To handle this, their approach adapted Apertium's dictionary format and integrated Constraint Grammar (CG) rules to handle mutation disambiguation before statistical analysis.

Building on similar principles, Tyers (2009) presented work on developing a Breton morphological analyser and part-of-speech tagger within the same Apertium framework. Like Welsh, Breton is a Celtic language with complex morphophonological alternations, including mutation, which pose difficulties for statistical models trained on limited data. Tyers demonstrated that a rule-based approach integrating Constraint Grammar (CG) could effectively handle these challenges, showing that CG rules offer a flexible and transparent way to encode language-specific grammatical constraints in low-resource contexts. This work reinforced the broader applicability of rule-based and CG-driven strategies across Celtic languages, motivating similar design choices in Welsh NLP systems.

2.2. Taggers and Toolkits

The part-of-speech tagger CyTag2 was introduced in (Neale et al., 2018) as part of the CorCenCC project (Knight et al., 2020b). It is a rule-based tagger leveraging lexical resources like dictionaries and gazetteers alongside Constraint Grammar rules to tokenize, lemmatize and tag parts-of-speech and mutation tags. Although state-of-the-art technology for POS taggers at the time were statistical models, the authors argued the limited availability of annotated data in the context of a low-resource language like Welsh did not allow for the development of this type of tagger. Additionally, they explained why this rule-based approach was much more suitable in dealing with Welsh due to its intricate morphology and mutation system.

Another important contribution is the Welsh Natural Language Toolkit (Cunliffe et al., 2022), which provides tokenisation, lemmatisation, POS tagging for Java-based systems. Their approach involved adapting the GATE (General Architecture for Text Engineering) framework for Welsh. Like earlier systems, it integrates external lexical resources but does not offer deeper grammatical interpretability beyond standard tagging.

2.3. Corpora

Several corpora have been developed to support Welsh NLP, though none approaches the scale and magnitude of the vast resources and datasets available for English. This is expected for a minoritised language, but it highlights the ongoing need within the Welsh NLP community for openly available, diverse, labelled, and carefully annotated datasets verified by proficient Welsh speakers.

One of the most significant projects to date is CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes, Knight et al., 2020a), launched in 2020. This project contains around 11 million tokens of naturally occurring Welsh, drawn from both audio and textual sources from diverse mediums including journals, TV programmes, as well as conversations which Welsh speakers could contribute directly via a crowdsourcing app specially designed for the project. Alongside the corpus, the project developed tools to help process it to support further

[1] Code available at https://github.com/ngutierrezrolon/welsh-mutation-trigger-labelling

downstream applications, such as the rule-based part-of-speech tagger CyTag2, the semantic tagger CySemTagger, and word frequency lists '*Yr Amliadur*'(Knight et al., 2020c).

Aside from the CorCenCC corpus, many others exist, such as parallel corpus of National Assembly proceedings (Donnelly, 2013) and the Siarad corpus (Deuchar et al., 2018). While these corpora provide valuable coverage, the availability of processed and labelled datasets remains limited. The creation of further benchmark datasets and similar resources, especially those designed natively in Welsh for Welsh, remains an urgent need. High quality resources are essential for the training of machine learning models, and fully automated pipelines risk propagating errors downstream. Even for phenomena like mutations, developing such a corpus is tricky when none of the tagging systems at the moment are reliable enough. Current automatic taggers like CyTag2 are highly accurate, but certainly not error free. Resources like these need to be developed with the guidance and validation of linguistic experts. More broadly, resource development should be undertaken as part of community-driven efforts, ensuring Welsh NLP benefits both researchers and the wider public.

Despite these advances, no existing system is able to explicitly analyse or explain the grammatical triggers behind mutations. Current tools can detect mutated forms but stop short of interpreting why a mutation occurs, whether due to lexical, grammatical, or syntactic triggers. This represents a key gap in Welsh NLP and in computational modelling for Celtic morphology more broadly.

This paper therefore builds directly on this foundation by proposing the first computational framework for mutation trigger labelling. In doing so, it extends prior rule-based efforts toward a more interpretable layer of linguistic labelling and explanation, aligning computational approaches with pedagogical and corpus-linguistic goals for Welsh.

3. Unsupervised Models

Two main model strategies were tested: a rule-based system and large language models (LLMs). Because this task is newly defined and no labelled data currently exists, all experiments were performed in an unsupervised setting. Instead, the aim was to evaluate the feasibility of automatic mutation trigger labelling, comparing the linguistic precision and coverage of the rule-based approach with the reasoning and adaptability of general-purpose multilingual LLMs as initial baselines for this task.

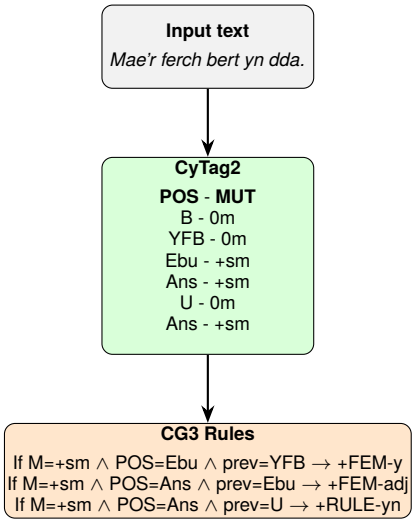

Figure 1: Rule-based pipeline for mutation trigger labelling using CyTag2 and CG3.

3.1. Rule-based Model

The rule-based model serves as a linguistically grounded baseline against which to evaluate more modern, data-driven approaches. It comprises of a pipeline (Fig. 1) using CyTag2 and a new set of CG rules to create an additional layer of tags on top of CyTag2's output to represent the mutation triggers if a mutation is detected by the tagger.

3.1.1. Mutation Identification

The first stage of the pipeline identifies mutations using CyTag2's tokenisation. This component consults dictionaries, lexicons and gazetteers[2] to determine whether each token exists in its observed form. Tokens not found in these resources are further analysed to account for conjugation, mutation, or non-standard spelling. Suspected mutations are "unmutated" by reversing possible mutation patterns and checking the resulting forms in the dictionaries. When a valid "unmutated" wordform is found, the token is relabelled with a mutation tag and the detected mutation pattern. For instance, the word *fara* would not be found in the dictionary, and variants *mara* (M+SM→F) and *bara* (B+SM→F) would be tested. Since *bara* matches a dictionary entry, the lemma would be *'bara'* and the tag `+sm` would be added to the output.

This process is largely language-agnostic, since it relies on dictionary membership and pattern matching rather than explicit grammar rules. Limitations include unrecognised words, coincidental matches with other valid words, surface-form overlap (e.g. *chi*, the aspirate mutated form of *ci* (dog),

[2]See (Neale et al., 2018) for a full description of the lexical resources used, including the Eurfa dictionary and CorCenCC project gazetteers.

Class	Family	Trigger	Example
Lexical	Prepositions (PREP)	o	Dw i'n dod **o Gaerdydd**.
	Possessives (POSS)	dy	Alun yw **dy dad** di.
	Simple (SIMP)	dyma	**Dyma daith** anhygoel.
	Prenominal Adjectives (PREADJ)	hen	Yr **hen destament**.
	Numeral (NUM)	Dau	**Dau gi** bach.
Morphosyntactic	Adverbial (ADF)	dydd	Dw i'n mynd nôl **ddydd** Llun.
	Feminine (FEM)	Y	**Y ferch** yw hi.
	Interpolation (INTP)	yma	Mae **yma ddigon** o fwyd.
	Rule (RULE)	Inflected object	Prynais i **fara**.

Table 1: Tag taxonomy families with examples.

has the same surface form as *chi* (you, formal/-plural)), and sub-word mutations. While improving mutation detection remains part of future work, the current implementation treats CyTag2's output as given, with the understanding that errors may propagate downstream.

3.1.2. Mutation Trigger Labelling

Once the mutations have been identified by the tagger, the next task is to determine their triggers. This is implemented through Constraint Grammar (VISLCG3), the same framework used by CyTag2 for morphological disambiguation.

We introduce a new taxonomy of tags dedicated specifically for different types of mutation triggers. Each mutated word receives a family tag and a specific trigger word tag, assigned based on its part-of-speech tag, the type of mutation identified, and its local context relative to potential triggers, following guidelines from *Y Treigladur* (Lewis, 2021) on both the specific lexical triggers featured in a word list, and the rules governing mutation section.

Note that this taxonomy is not strictly aligned with the morphosyntactic changes of a mutated word, but with the function words and syntactic structures that trigger mutation. While existing Welsh NLP tools such as CyTag2 (Neale et al., 2018) and UDPipe (Straka and Straková, 2017) provide valuable morphological and syntactic annotation, including disambiguating part-of-speech or identifying mutated word forms, they do not capture the causal mechanisms behind mutations. Our taxonomy is fundamentally different: it is not aligned with the mutated word itself, but with the function words or syntactic structures that trigger the mutation. In other words, we focus on why a mutation occurs, rather than merely detecting that it has occurred. Existing tools can label a mutated token, but are unable to indicate which preceding element is responsible, or whether that element is actively triggering the mutation. No existing system provides an explicit, interpretable analysis of lexical, morphological, or syntactic mutation triggers, leaving a gap in computational modelling of Welsh. By targeting triggers rather than mutated forms, our framework provides a layer of linguistic annotation that is both explanatory and pedagogically useful, complementing prior morphosyntactic analyses.

A total of **234 CG rules were written** to append **179 distinct tags**. Some tags are appended by multiple rules; for instance, the tag FEM-adj is assigned by two rules to account for feminine nouns which might be POS-tagged as either *Ebu* (feminine) or *Egbu* (ambiguous gender, reflecting regional variants), though the presence of a mutation is evidence that the element is feminine. In other cases, multiple rules were also used to capture different surface wordforms such as infixed pronouns (e.g. *'w = ei*) as the trigger.

The tags are organised into different families corresponding to major categories of mutation triggers, mostly separated by PoS, and distinguishing between lexical and morphosyntactic triggers, following (Mittendorf and Sadler, 2006; King, 1993; Ball and Müller, 1992). The full family taxonomy is presented in Table 1, showing each family and class with one particular trigger and a small example of both the trigger and mutation in context.

Because multiple trigger tags may apply to a single mutation, the grammar applies rules in a fixed order, ensuring that the most plausible triggers are considered first. All simple lexical triggers are applied at the start, since these cases require a short context window and are generally unambiguous. The more complex morphosyntactic tags are applied near the end, only after all the simple categories have been exhausted. This is particularly relevant for the case of direct object mutations (DOM). In Welsh, the direct object of an inflected verb undergoes soft mutation under certain circumstances. This is one of the most common syntac-

tic triggers, belonging to the RULE family. Note that Welsh syntax has a verb-subject-object (VSO) canonical order, which contributes to the difficulty of identifying if an noun phrase (NP) is the object. However, because CyTag2 does not include any syntactic information, the CG rules are limited. In practice the rule assigns DOM tags to tokens that are not instances of this mutation, simply because an inflected verb is likely to appear in the sentence. In order to minimise the amount of times this tag gets added incorrectly, the DOM rule is placed last, such that it functions as a fallback only when no other trigger is available.

Finally, mutations caused by vocative triggers or by interpolation cannot be fully captured. Part of the reason is that common interjections that trigger vocative mutations are not tagged correctly by CyTag2. Additionally, since Cytag2 has no references to syntactic structures like subject-object dependencies, it cannot identify interpolated structures. Only some common trigger words which occur in interpolation (such as *hefyd*, *yma*) have been added as tags for occurrences of interpolation, but the structure itself often occurs with no lexical elements indicating it.

3.2. In-Context Learning with LLMs

The LLM-based experiments were conducted in a few-shot prompting setting. The two models chosen were GPT-4o-mini and Claude-Sonnet-4, since they represent the current state-of-the-art for multilingual large language models. The LLMs had an in-context learning approach by prompting both models with examples and explanations of the labels expected.

Each input sentence contained the target mutated word in uppercase, and prompts explicitly defined the required output format and label taxonomy to guide the models' responses, explaining shortly the grammatical families, without explicitly naming every label for brevity but keeping at least one example of a potential label for each family. The models are also provided with a fall-back strategy that if the trigger does not seem to fit any of those categories, to default to the Simple family. The inclusion of this structure, along with example sentences in the prompt, ensured the models had guidance on the task while remaining generalisable to unseen data. The full prompt used is included in Figure 2.

```
You are a Welsh mutation detector and
    analyser.
For each sentence, the TARGET word
    will be in UPPERCASE.
Return ONLY the correct mutation type
    and reason in the following
    format:
    +<mutation>, <reason>

### Mutation Types:
+sm = soft mutation
+nm = nasal mutation
+am = aspirate mutation
+hm = h-prothesis

### Tag Families:
- SIMP-(trigger) -> simple triggers
    (default if unsure)
- PREADJ-(trigger) -> pre-nominal
    adjectives
- POSS-(trigger) -> possessives
- PREP-(trigger) -> prepositions
- NUM-(trigger) -> numerals
- FEM-(trigger) -> feminine triggers
    (y, adj, num, etc.)
- ADF-(trigger) ->
    adverbs/determiners (bob, ddim,
    rhy, etc.)
- RULE-(trigger) -> grammatical
    triggers (ydau, ydwy, ni, a,
    rhaid, yn, etc.)
- INTP-(trigger) -> syntactic
    interpolation (yma, hefyd, etc.)

If the trigger doesn't fit any
    special class, use SIMP-(trigger).
The trigger should always be in
    lowercase, except for SIMP-A.

### Examples:
Sentence: Sut mae fy NHAD i?
Response: +nm, POSS-fy

Sentence: Beth DDIGWYDDOD?
Response: +sm, RULE-a

Sentence: Mae'r ferch DDA yn canu.
Response: +sm, FEM-adj

Your new sentence is:
```

Figure 2: Full prompt used for Claude and GPT classification.

4. Experimental Setting

4.1. Datasets

The evaluation dataset was compiled from three primary sources: D.A. Thorne's *A Comprehensive Welsh Grammar* (Thorne, 1993) and *Taclo'r Treigladau* (Thorne, 1997), and the CorCenCC corpus (Knight et al., 2020b).

4.1.1. Grammar Books

Both grammar references by D.A. Thorne were chosen for their reliability, such that each mutation is determined by a Welsh language expert. The textbooks contain subsections about mutation, carefully labelled by each trigger, and including examples which are extracts from literary and journalistic Welsh. These are used as the gold labels for this section of the dataset. This source was chosen to ensure grammatical correctness and adherence to the standard written form.

Thorne notes in his grammar guides that some of the triggers listed are optional or stylistic. For instance, post-nominal adjectives which are used pre-nominally cause mutation, though this is more a literary device. Such variation reflects the natural diversity of Welsh usage across text types.

4.1.2. Corpora

The CorCenCC corpus contributed examples from colloquial real language use. Because these examples are not expert-annotated, only simple lexical triggers were included to preserve accuracy, since this type of mutation always occurs on the following word, so it can be assumed that it is the main trigger causing the mutation which follows. Examples from the corpora were obtained using the KWIC (Key Word in Context) tool,[3] searching for the specific trigger words that always cause lexical mutation, and checking that the following word is tagged with the correct type of mutation.

Examples from CorCenCC were selected filtering by trigger word, part of speech tag, mutation type, and metadata such as register or region. This allowed inclusion of spoken and colloquial varieties specifically. This approach was also useful for triggers with multiple functions, like the wordform *a* which could be a predicative particle, an interrogative particle or a conjunction, which cause different types of mutation. Additionally, certain triggers exhibit colloquial variants, like the first-person singular possessive pronoun *fy*, which triggers nasal mutation, and in colloquial varieties can appear as *'y* or even just *y*, potentially causing ambiguity due to its similarity to the definite article *y*.

The dataset was designed not to reflect natural frequency of triggers, but to test coverage for a wide range of triggers, including some uncommon and literary ones. The evaluation dataset is imbalanced, with simple triggers forming the largest class. Figure 3 shows the distribution of examples according to lexical or morphosyntactic subclass.

[3] https://corpus.corcencc.org/

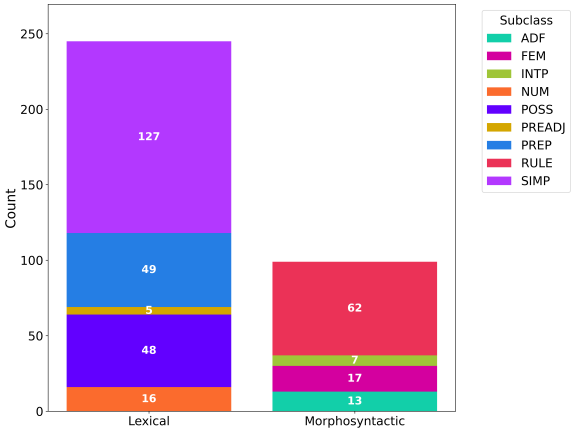

Figure 3: Distribution of examples by trigger family and class

4.2. Metrics

To assess system performance in detecting, identifying and labelling Welsh mutations, four standard metrics were used: **accuracy**, **precision**, **recall**, and **F1-score**. These capture complementary aspects of performance: accuracy provides the overall success rate, precision measures reliability of predicted triggers, recall reflects overall coverage, and F1 balances precision and recall. Conducting the analysis at the subclass level also reveals insights into the behavior of specific grammatical families.

Since trigger labelling is formulated as a multiclass task, the computation of precision and recall for the rule-based model was adapted as follows:

- **True Positive:** the correct mutation trigger was predicted;
- **False Positive:** an incorrect mutation trigger was predicted;
- **False Negative:** no prediction was made, either because the word was not found or no mutation was detected;
- True negatives are not applicable since the dataset only includes instances containing mutations.

However, these metrics cannot apply to the LLMs, since they always produce a prediction. Since standard classification metrics do not capture cases where LLMs produce partially correct outputs, three complementary evaluation measures were introduced for evaluating this type of model:

- **Full-match accuracy:** both the mutation family and the trigger word are correctly predicted;
- **Family-match accuracy:** the correct mutation family is predicted, regardless of the trigger word;

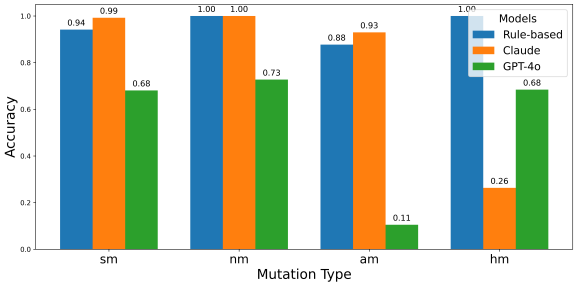

Figure 4: Accuracy of Mutation Detection across Models

- **Trigger-match accuracy:** the correct trigger word is predicted, regardless of the family label.

5. Results and Discussion

5.1. Mutation Detection

The rule-based model, which integrates CyTag2 for mutation detection, achieved a strong 94% accuracy in mutation detection. Among the LLMs, Claude matched the rule-based model, while GPT4o-mini achieved 59%.

Figure 4 presents accuracy by mutation type. Despite relying on heuristic rules and dictionary-based identification, CyTag2 performs consistently across mutation types: 94% for Soft Mutation (SM), 100% for Nasal Mutation (NM), 88% for Aspirate Mutation (AM), and 100% for H-Mutation (HM). NM and HM, which are highly regular and exclusively lexically triggered, showed perfect accuracy, while SM and AM proved more challenging due to overlapping mutated forms and ambiguous contexts.

In contrast, the LLMs exhibited greater variation across mutation types. GPT struggled particularly with aspirate mutation (11%) but performed moderately on NM (73%), SM (68%), and HM (68%). Claude achieved near-perfect results on SM (99%), NM (100%), and AM (93%), but only 26% on HM, with a tendency for confusing it with AM. Overall, Claude matched the rule-based model, while GPT lagged significantly, suggesting greater robustness in Claude's Welsh training data.

5.2. Mutation Trigger Labelling

Trigger prediction proved much more difficult overall. The rule-based model reached 65% overall accuracy, with precision of 0.831, Recall of 0.742, and F1-score of 0.784, indicating that failure to predict a trigger is more frequent than wrong guesses. Figure 5 shows subclass performance: lexical triggers are labelled accurately, while morphosyntactic triggers exhibit a much weaker performance and prove considerably more difficult to label reliably.

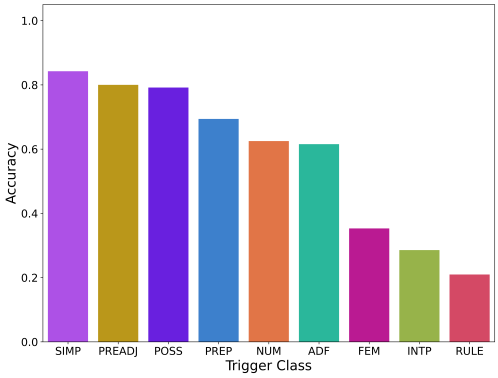

Figure 5: Accuracy of rule-based model for trigger labelling across families

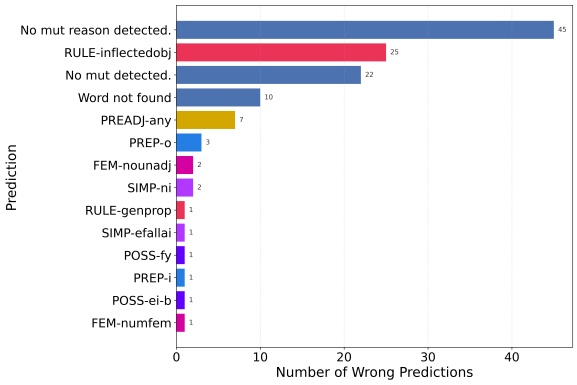

Figure 6: Distribution of Errors of rule-based model. FN errors in light-blue, all other are FP errors.

Manual analysis of FP and FN errors was conducted on a sample, as shown in Figure 6. Most FP errors (56%) came from over-assigning the tag 'RULE-inflectedobj', whenever an inflected verb occurs. Without syntactic dependency data, the system is unable to determine if the mutated word is actually the verb's object. FN errors were dominated by 'no reason detected' (58%), 'no mutation detected' (29%), and 'word not found' (13%). This pattern reflects strong precision but limited recall, reflecting the challenge of handling complex morphosyntactic triggers within CG's framework.

For the LLMs, accuracy on trigger labelling was substantially lower than the rule-based model. GPT-4o-mini achieved 18% full-match accuracy, while Claude reached 36%. Both models performed better on partial matches, due to the complexity of multi-label reasoning.

GPT typically recognised the general mutation family but failed to identify the precised trigger, while Claude outperformed it on all metrics (41% full, 52% family, 56% trigger). The rule-based system, while limited by coverage, remained the most reliable overall, showing higher precision and inter-

Model	Match	POSS	PREP	SIMP	NUM	PREADJ	INTP	FEM	ADF	RULE
Rule	–	**79.2**	**69.4**	**84.3**	62.5	**80.0**	**28.6**	35.3	61.5	21.0
GPT	Full	29.2	8.2	17.3	37.5	0.0	14.3	29.4	46.2	4.8
	Family	35.4	34.7	37.0	56.2	0.0	**28.6**	29.4	46.2	27.4
	Trigger	39.6	22.4	40.2	37.5	20.0	14.3	35.3	53.8	8.1
Claude	Full	66.7	59.2	22.8	68.8	20.0	**28.6**	35.3	61.5	9.7
	Family	77.1	71.4	29.9	68.8	40.0	**28.6**	**41.2**	61.5	**48.4**
	Trigger	72.9	69.4	67.7	**81.2**	60.0	**28.6**	**41.2**	**69.2**	9.7

Table 2: Per-family accuracies for Rule-based, GPT, and Claude models, split by Full, Family, and Trigger matchings. Bold indicates the best result per column.

pretability.

In summary, all models handled lexical triggers best. The rule-based model and Claude performed comparably in mutation detection, but the rule-based approach far outperformed both LLMs in mutation trigger labelling, reflecting its explicit linguistic structure and reliability.

5.3. Class-Level Analysis

Table 2 summarises per-family accuracies for all models, including Full, Family, and Trigger matchings for the LLMs. Across models, lexical triggers were reliably identified, while morphosyntactic triggers show consistently lower performance. Overall, the rule-based model maintains the highest accuracy for most families.

For lexical families, the rule-based model consistently achieves the highest accuracy across POSS, PREP, SIMP, and PREADJ, reflecting its explicit encoding of lexical triggers. GPT-4o-mini struggles across these categories, showing particularly low Full-match performance, though Family-match scores indicate that it often correctly identifies the grammatical category even when the exact trigger is missed. Claude performs slightly better than GPT for these lexical families, but the rule-based model remains superior for all families except NUM, where Claude's Full and Trigger matches surpass the rule-based model.

The FEM family presents a mixed pattern: the rule-based model leads in Full-match accuracy, but Claude slightly outperforms it in Trigger-match, indicating that Claude can detect mutations triggered by feminine nouns or articles but is less precise in full multi-label assignment. GPT shows modest performance across all matching types.

For other morphosyntactic families, such as INPT, RULE, and ADF, LLMs show higher Family-match scores than Full or Trigger matches. Both GPT and Claude achieve Family-match of 28.6% for INTP, Claude reaches 48.4% for RULE family matching and 69.2% in ADF Trigger matching, surpassing the rule-based model. This pattern demonstrates that LLMs can recognise the presence of complex, context-dependent morphosyntactic triggers but struggle to identify their exact forms or assign all associated labels correctly. Full and Trigger matches remain low, reflecting the challenge of multi-label prediction.

Finally, across all models, lexical triggers are detected reliably, with rule-based approaches providing the most consistent results. LLMs show potential for certain morphosyntactic triggers, but struggle to achieve high Full-match accuracy. These patterns indicate that explicit linguistic rules remain advantageous, though LLMs show potential in identifying certain context-dependent phenomena.

6. Qualitative Analysis

This section analyses common errors in both the rule-based system and the LLM baselines, focusing on causes, linguistic patterns, and implications for future work. A randomised sample of 100 instances of errors were manually analysed, and for the rule-based model the CyTag2 output was also manually analysed to determine where the error occurred.

6.1. Rule-based Model errors

Errors in the rule-based model stem from two main sources:

6.1.1. CyTag2 errors

53% of errors originate from upstream issues with CyTag2, and correspond to the False Negative (FN) errors. Manual analysis of the tagging output shows that incorrect tokenisation was applied for some apostrophes, particularly forms like *i'w* were not split correctly, obscuring pronouns and preventing trigger recognition. Set phrases such as *wrth gwrs* (of course) were compounded by CyTag2 into wrth_gwrs, which hid the mutated element from the system. Additionally, certain common function words like *ni* or *fe* were mistagged as pronouns rather than particles, blocking the rule application.

Finally, ambiguous words like *deg* (fair +SM) vs (ten +no mutation) caused false negatives.

6.1.2. Rule errors

46% were Rule errors, corresponding to the False Positive (FP) errors. Overly specific or narrow constraints caused misses. For instance *RULE-ni* required *ddim* to appear, ignoring other possible negative contexts, or in some cases, no lexical trigger at all to indicate the rule, case in which the aspirate mutation is the only indicator of negation. Additionally, in some cases ambiguity between syntactic roles (e.g. *Cafodd Llwyd lwyddiant nodedig*, where a proper noun followed by a mutated noun let the system assume a genitive structure) caused misclassification of genitive as opposed to object triggers. Finally, the limitations of Constraint Grammar implicated some triggers could not be encoded, particularly equatives, parenthetic verbs, or interpolations.

Overall, these issues stem from limitations in CyTag2's surface-level handling and CG's lack of deeper syntactic structural representations.

6.2. LLM Errors

The GPT-4o-mini and Claude models fail for different reasons. GPT tends to over-generalise: it frequently assigns *SIMP-* family tags to nearby words, likely due to over-adherence to the prompt, or produces labels based on other elements in the sentence, unrelated to the specific mutation. It often identifies the grammatical family but not the correct trigger word, particularly failing to distinguish between *SIMP* and *RULE* classes. On the other hand, Claude performs better but still shows systematic family confusion. Similarly to GPT, it occasionally defaults to *SIMP* triggers or attaches to triggers in the sentence which are unrelated to the actual target mutation.

This qualitative analysis highlights some of the key differences between the rule-based system and the comparison models. The rule-based system is precise and interpretable but sensitive to downstream tagging errors and unable to distinguish syntactic dependencies. In contrast, the LLMs fail for different reasons, often over-generalising based on the prompt.

7. Conclusion and Future Work

This paper introduced and explored a new research direction in Welsh NLP: the task of mutation trigger labelling, of determining the grammatical or lexical triggers that cause initial consonant mutations in context. While mutation has been described and discussed extensively in Welsh linguistics, this is, to our knowledge, the first computational attempt to formalise and operationalise the question of why mutations occur.

Two complementary approaches were explored to test the feasibility of the task: a linguistically informed, rule-based system that integrates Constraint Grammar rules, and a set of large language models prompted in few-shot setting. Their comparison demonstrates the challenges and opportunities in mutation trigger labelling. While the rule-based model remains more robust and interpretable, the exploratory LLM results show potential for broader generalisation once sufficient data becomes available. Together, these experiments validate that tagging mutation triggers can be approached computationally, and establish benchmarks for future research.

One potential way of addressing the limitations of the CG approach would be the integration of syntactic information derived from UDPipe models, which offer dependency parsing for Welsh. Incorporating such information could be used to improve the detection of morphosyntactic mutation triggers.

Beyond model performance, the main contribution of this work is reframing Welsh mutation as a complex, linguistic computational task, opening a new research space for explainable and interpretable grammatical analysis within a low-resource setting in Welsh.

In summary, this paper presents a pioneering step toward interpretable Welsh NLP. By defining the task of mutation trigger labelling and evaluating early approaches, it sets out a research agenda for future systems capable of explaining linguistic structure in transparent, context-aware ways. This work contributes not only to the digital development of Welsh, but to the broader field of explainable NLP for the benefit of low-resource and morphologically rich languages.

Limitations

Our approach relies on Constraint Grammar (CG) rules and the CyTag2 pipeline, which introduces several limitations. CG-based rules are inherently unable to detect certain syntactic mutation triggers, such as the direct object of an inflected verb or interpolated syntactic structures. Some triggers are also subject to linguistic debate, as noted by D.A. Thorne on the variability of certain mutation triggers, and some that are context-specific like literary mutations, making their formalisation ambiguous. Reliance on CyTag2 further constrains the approach, since its errors necessarily propagate downstream. If the tagger fails to detect a mutation or mislabels the part-of-speech, our pipeline cannot recover or correctly annotate the feature. A broader limitation is the scarcity of fully labelled corpora for Welsh, which necessitates the use of

unsupervised or semi-automatic methods, limiting coverage and evaluation. These factors collectively mean that, while our tool can correctly label many mutation triggers, it does not capture the full spectrum of possible contexts in contemporary Welsh usage. Additionally, a further limitation is that some datasets used, such as the Thorne grammar examples, are under copyright and cannot be publicly redistributed, limiting the reproducibility and external evaluation of the mutation detection rules.

Ethics Statement

The work presented engages with Welsh language norms and usage, which raises some ethical considerations. Our dataset is derived largely from grammar books and literary sources, reflecting a prescriptive standard and under-representing colloquial and dialectal varieties. Computational approaches that encode standardised language rules risk marginalising these equally valid forms of the language. Additionally, mutation itself is variable in modern usage and its application differs across speakers, with linguists debating whether traditional mutation rules remain fully applicable. Our tool is designed as a descriptive instrument to label mutation triggers, rather than enforce prescriptive norms beyond appropriate contexts. It can support analysis of dialectal and colloquial variation, but caution is warranted to avoid misrepresenting the diversity of contemporary Welsh language use. Computational linguistics should aim to reflect language use descriptively rather than impose prescriptive standards.

8. Bibliographical References

G. M. Awbery. 1986. Moves towards a simpler, binary mutation system. In H. Andersen, editor, *Sandhi phenomena in the languages of Europe*, volume 33 of *Trends in Linguistics, Studies and Monographs*, pages 161–166. Mouton de Gruyter, Berlin.

M. J. Ball. 1984. *Sociolinguistic aspects of the Welsh mutation system*. Ph.D. thesis, University of Wales. Unpublished doctoral dissertation.

Martin J. Ball and Nicole Müller. 1992. *Mutation in Welsh*. Routledge, London / New York.

N. Coupland and M. J. Ball. 1989. Welsh and english in contemporary wales: Sociolinguistic issues. *Contemporary Wales*, 3:7–40.

Daniel Cunliffe, Andreas Vlachidis, Daniel Williams, and Douglas Tudhope. 2022. Natural language processing for under-resourced languages: Developing a welsh natural language toolkit. *Computer Speech & Language*, 72:101311.

John Davies. 2014. *The History of the Welsh Language*. University of Wales Press, Cardiff.

Margaret Deuchar, Peredur Webb-Davies, and Kevin Donnelly. 2018. *Building and Using the Siarad Corpus: Bilingual Conversations in Welsh and English*, volume 81 of *Studies in Corpus Linguistics*. John Benjamins.

Kevin Donnelly. 2013. Kynulliad3: A corpus of 350,000 aligned Welsh and English sentences from the third assembly (2007–2011) of the National Assembly for Wales. http://cymraeg.org.uk/kynulliad3.

Gareth King. 1993. *Modern Welsh: A Comprehensive Grammar*. Routledge, London.

Dawn Knight, Steven Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, Enlli Môn Thomas, Alex Lovell, Jonathan Morris, Jeremy Evas, Mark Stonelake, Laura Arman, Joshua Davies, Ignatius Ezeani, Steven Neale, Jennifer Needs, Scott Piao, Mair Rees, Gareth Watkins, Lowri Williams, Vignesh Muralidaran, Bethan Tovey-Walsh, Laurence Anthony, Tom Cobb, Margaret Deuchar, Kevin Donnelly, Michael McCarthy, and Kevin Scannell. 2020a. Corcencc: Corpws cenedlaethol cymraeg cyfoes – the national corpus of contemporary welsh. *Cardiff University*.

Dawn Knight, Steven Morris, Tony Fitzpatrick, Paul Rayson, Irena Spasić, and Enlli M. Thomas. 2020b. The national corpus of contemporary welsh: Project report | y corpws cenedlaethol cymraeg cyfoes: Adroddiad y prosiect. ArXiv preprint arXiv:2010.05542.

Dawn Knight, Stuart Morris, Bethan Tovey-Walsh, Tom Fitzpatrick, and Laura Anthony. 2020c. Yr amliadur: Frequency lists for contemporary welsh. Technical report, Cardiff University.

D. Geraint Lewis. 2021. *Y Treigladur - A Check-List of Welsh Mutations (Argraffiad Newydd)*. Y Lolfa.

Ingo Mittendorf and Louisa Sadler. 2006. A treatment of welsh initial mutation. In *Proceedings of the LFG '06 Conference*, pages 343–364, Stanford, CA. CSLI Publications.

Steve Neale, Kevin Donnelly, Gareth Watkins, and Dawn Knight. 2018. Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in welsh. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, volume 3, pages 168–173, Miyazaki, Japan. European Language Resources Association (ELRA).

Office for National Statistics. 2011. 2011 census for wales. https://www.ons.gov.uk/census/2011census. Accessed 13 October 2025.

Milan Straka and Jana Straková. 2017. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphosyntactic tagging, lemmatization and dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

David A. Thorne. 1993. *A Comprehensive Welsh Grammar*. Blackwell, Oxford.

David A. Thorne. 1997. *Taclo'r Treigladau*. Gwasg Gomer, Llandysul.

Francis M. Tyers. 2009. Rule-based augmentation of training data in Breton-French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Francis M. Tyers and Kevin Donnelly. 2009. Apertium-cy: A collaboratively-developed free rbmt system for welsh to english. *The Prague Bulletin of Mathematical Linguistics*.

Welsh Government. 2011. Welsh language measure 2011. https://www.legislation.gov.uk/mwa/2011/1/contents. Accessed 13 October 2025.

Welsh Government. 2017. Cymraeg 2050: A million welsh speakers. https://www.gov.wales/cymraeg-2050-welsh-language-strategy. Accessed 13 October 2025.

Welsh Government. 2018. Welsh language technology action plan. https://www.gov.wales/welsh-language-technology-action-plan. Accessed 13 October 2025.