Open-Source Thesaurus Development for Under-Resourced Languages: a Welsh Case Study

Nouran Khallaf[1], **Elin Arfon**[2], **Mo El-Haj**[1], **Jonathan Morris**[2], **Dawn Knight**[2], **Paul Rayson**[1],

[1]Lancaster University, [2]Cardiff University
{n.khallaf1,m.el-haj,p.rayson}@lancaster.ac.uk,
{ArfonE,morrisj17,knightd5}@cardiff.ac.uk

Tymaa Hammouda[3] **and Mustafa Jarrar**[3]

[3]Birzeit University
{1171779@student.birzeit.edu,mjarrar@birzeit.edu}

Abstract

This paper introduces an open-access, user-friendly online thesaurus for the Welsh language, aimed at enriching digital resources for Welsh speakers and learners. Utilising advances in Natural Language Processing (NLP), our approach combines pre-existing word embeddings, a Welsh semantic tagger, and human evaluation to establish related terms. In this case, an initial list of 250 words was expanded by adding 6,953 synonyms provided by linguists, creating a more extensive foundation for building the gold-standards. With this expanded list, when a user queries a particular word, the thesaurus presents all of its synonyms, allowing them to choose from a wider range of options. This is especially helpful when a user is unsure of the exact word they want to use or wants to explore different ways to express a concept. The resulting thesaurus offers a comprehensive, reliable resource for Welsh language users, fostering enhanced communication and expression. Our work promotes Welsh NLP and showcases NLP's potential to support under-resourced languages. The thesaurus will be accessible via a bilingual website, and the accompanying Python code will be available in a bilingual, public GitHub repository, and it will be available as a web service. Our approach presents a more efficient, cost-effective method for thesaurus creation, with potential applicability to other under-resourced languages.

1 Introduction

The Welsh language is a critical component of Welsh cultural identity and heritage. The latest (2021) census reports that 538,300 people aged three and over consider themselves to be speakers of the language, which corresponds to 17.8% of the population[1]. Despite its importance, Welsh language users face significant challenges in accessing digital resources, particularly when it comes to reference tools such as a thesaurus. This is a significant barrier to the promotion and preservation of the Welsh language, as it limits the ability of users to effectively communicate and express themselves in Welsh. While there are some Welsh language thesauri currently available, such The Gweiadur project[2], which is still in beta, these resources are limited in scope and do not provide the level of functionality that users need to fully utilise the Welsh language. As such, the development of a comprehensive Welsh language thesaurus is essential for the promotion and preservation of the Welsh language, and to enable Welsh language users to communicate effectively and express themselves in their native language.

Currently, the creation of a comprehensive Welsh language thesaurus involves significant manual effort, with lexicographers and linguists required to curate the content and ensure its accuracy. This process is time-consuming, expensive, and often reliant on the availability of skilled professionals. By leveraging recent developments in NLP and word embeddings, we can create a thesaurus for Welsh that is faster, more cost-effective, and more scalable. Word embeddings provide a way to identify and group words based on their meaning and usage, allowing for the automated creation of a network of related words. This significantly reduces the need for human intervention, enabling us to create a comprehensive Welsh language thesaurus that can be easily updated and maintained over time. In this way, our approach has the potential to significantly enhance the availability of digital resources for Welsh language users, facilitating effective communication and expression in Welsh.

This paper presents the development of an open-access, freely available online thesaurus for the Welsh language, which aims to enhance digital resources available to Welsh speakers and learners

[1] Welsh Language in Wales (Census 2021) https://www.gov.wales/welsh-language-wales-census-2021-html

[2] https://www.gweiadur.com/thesaurus

(financed by Welsh Government). Our approach leverages recent advances in NLP, using preexisting word embeddings to identify related words, a Welsh semantic tagger (Piao et al., 2017) and human evaluators to refine the similarities. This innovative methodology has shown success with more widely spoken languages, such as French (Hazem and Daille, 2018), and our work represents an important contribution to under-resourced languages such as Welsh, where the availability of digital resources is limited. The resulting thesaurus provides a comprehensive and reliable resource for Welsh language users, enabling more effective communication and expression in Welsh. In addition, our methodology has the potential to be applied to other under-resourced languages, offering a more automated and cost-effective approach to thesaurus compilation. This paper contributes to the advancement of Welsh language NLP and demonstrates the potential for NLP methods to benefit under-resourced languages.

Recent developments in NLP have enabled the creation of word embeddings, which involve transforming words in a corpus (collection of speech) to vectors. Words that are similar in meaning or association are mapped to a similar location in the vector space, allowing for the identification of related words and the creation of a network of related words. For the language user, this represents a valuable resource that goes beyond traditional thesauri, as it enables them to discover and explore a wider range of related words and concepts.

In our project, we used pre-existing word embeddings for Welsh (Corcoran et al., 2021) to find similar words, providing a starting point for the development of our Welsh language thesaurus. However, to ensure the accuracy and relevance of the thesaurus, we further refined the similarities using the Welsh Semantic Tagger, which helped to ensure that similar words belong to the same Part-of-Speech (POS) and the same semantic field as the original. This process will enable us to create a comprehensive and reliable resource for Welsh language users.

The resulting thesaurus will be publicly available as a fully bilingual and user-friendly website. Additionally, the accompanying python code will be available through a bilingual, public-facing GitHub repository, enabling other researchers to build on our work and further improve Welsh language NLP. In this way, our work will contribute to the advancement of Welsh language NLP and provide an additional valuable resource for Welsh language users (El-Haj et al., 2022a,b; Ezeani et al., 2022; Morris et al., 2022).

2 Related Work

2.1 Low Resourced Languages: The Importance of Welsh Language and Technology

Welsh is an official language in Wales and current legislation places responsibilities on certain public bodies to provide bilingual services, including digital resources[3]. However, the availability of such resources for Welsh language users arguably remains limited, particularly when it comes to reference tools such as a thesaurus.

The Welsh government has made efforts to safeguard and promote the use of the Welsh language (Carlin and Chríost, 2016), but the uptake of Welsh language websites and e-services remains relatively low (Cunliffe et al., 2013). One reason for this may be the assumption that the language used in such resources will be too complicated. However, guidelines exist for creating easy-to-read documents in Welsh, including the use of everyday words rather than specialised terminology and a neutral register (Arthur and Williams, 2019; Williams, 1999).

The work presented in this paper aims to contribute to the digital infrastructure of the Welsh language, by developing an open-access, freely available online thesaurus for Welsh speakers and learners alike, including the introduction of Welsh Language Standards which place requirements on public institutions to provide fully bilingual web content (Carlin and Chríost, 2016).

The resulting thesaurus will complement the suite of Welsh language technologies, making it easier for content creators and Welsh readers to communicate effectively in Welsh. Additionally, the thesaurus will be of use to Welsh-medium educators and learners, who can use it as a pedagogical tool to better understand the nuances of the Welsh language. In addition, the work contributes to the advancement of Welsh language NLP and demonstrates the potential for NLP methods to benefit under-resourced languages. By leveraging the power of technology, we can help make the

[3] Welsh Language Standards www.welshlanguagecommissioner.wales/public-organisations/welsh-language-standards

Welsh language more accessible and easier to use for Welsh speakers and learners alike.

2.2 Semantic Field Annotation

In terms of thesaurus compilation for low resourced languages, we can benefit from linguistic knowledge already embedded in any existing taxonomies or ontologies if they are available, and in the case of Welsh, one such key resource is the UCREL Semantic Analysis System (USAS)[4]. Originally developed for English text (Rayson et al., 2004), a similar system was subsequently created for Welsh during the CorCenCC project[5] (Piao et al., 2018). USAS is a knowledge based annotation system, drawing on lexicons of single words and multiword expressions (MWEs) that have been manually created or checked by native speakers, to provide lists of potential coarse-grained word senses for each word or MWE. The USAS tagger then uses a variety of disambiguation methods to select the most likely meaning in context, employing a set of 232 semantic fields for its labelling of semantic tags or concepts[6]. For Welsh, the tagger achieves coverage of 91.78% in text, thus providing a wide set of information linking words to others that share the same conceptual category, in this case, via the semantic field tagging.

2.3 Thesaurus Creation

Creating a thesaurus involves compiling a list of related terms organised by the meaning of the words. There are several methods for creating a thesaurus, including manual and automated methods.

Manual methods involve human experts compiling lists of related terms based on their knowledge of the subject area. These experts may use a variety of sources, such as domain-specific dictionaries, thesauri, and other reference materials to identify related terms. This method is time-consuming but can produce high-quality thesauri (Aitchison et al., 2000).

Automated methods use either statistical algorithms or NLP techniques to identify relationships between words. This method depends on using large corpora to identify related terms based on their co-occurrence patterns in the corpus. This method is faster than manual methods but the results can be less accurate (Manning et al., 2008). There are several NLP approaches that can be used for the creation of thesauri. Distributional semantics, semantic clustering, semantic role labelling, graph-based algorithms, and other techniques such as Latent Semantic Analysis (LSA) (Turney, 2007), Latent Dirichlet Allocation (LDA), and word embeddings are all effective methods for identifying relationships between words and grouping them based on their semantic meaning.

One example of an NLP approach to thesaurus creation is the use of distributional semantics, which models the meaning of a word based on the distribution of its context words in a large corpus. This approach has been used to create a variety of thesauri in different languages, including English (Turney, 2007). Another semantic clustering algorithm that group words together based on their semantic similarity. As such, the WordNet thesaurus was created using this method, where words are organised into synsets (sets of synonyms) based on their meanings (Fellbaum, 1998). Semantic Role Labelling (SRL), which identifies the roles that words play in a sentence is another method for word grouping. For example, the WordNet Domains thesaurus was created using SRL, where the roles played by nouns in a corpus of texts were used to identify the semantic domains of the words (Magnini et al., 2000).

Hybrid methods combine manual and automated methods, using human experts to validate the results of automated algorithms. This method can produce high-quality, more efficient and cost-effective thesauri than relying solely on manual methods. Nonetheless, the use of NLP techniques for thesaurus creation has shown promise in creating comprehensive and accurate thesauri.

Latest NLP techniques that have been used for thesaurus creation include Word Embeddings. Landthaler et al. (2018) proposed a method for extending existing thesauri by leveraging word embeddings and the intersection method. Their approach involved using word embeddings to identify candidate synonyms for each entry in an existing thesaurus, and then intersecting these candidates with the existing synonym sets to identify and validate new synonyms. The authors evaluated their method on an existing thesaurus of human resources management terms and demonstrated that their method significantly improved the coverage and precision of the thesaurus, while maintaining its consistency and coherence.

[4] https://ucrel.lancs.ac.uk/usas/

[5] https://corcencc.org/

[6] https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

Our approach utilises recent developments in NLP to identify related words by using pre-existing Welsh word embeddings. We further refine these similarities through a Welsh semantic tagger and human evaluators to create a reliable and comprehensive resource for users of the Welsh language. Our method has been tested on existing dictionaries and graph-based thesauri, and is described in detail in the rest of the paper.

3 Words lists description

In order to build and evaluate our thesaurus for Welsh, we began by creating gold-standard synonyms for a list of 250 words. This list was comprised of 84 NOUN lemmas, 84 VERB lemmas (excluding conjugated verbs), and 82 ADJECTIVE lemmas, all taken from a frequency list of Welsh words (Knight et al., 2020).

We started by obtaining a list of 500 most frequent Welsh words from the Welsh National Corpus (Knight, 2020; Knight et al., 2021), specifically from the Yr-Amliadur.pdf document available on the CorCenCC website (Knight et al., 2020). From this list, we selected roughly equal numbers of nouns, adjectives, and verbs, excluding any duplicates or conjugated verbs.

To ensure a diverse selection of words for our gold standard, we included items from both the beginning and final parts of the list. We also included a number of homophones. This approach allowed us to capture a range of word types and usage contexts, including less common words that may be important for Welsh language users but are not frequently encountered in everyday language.

Our aim in building this gold-standard was to provide a reliable set of synonyms that we could use to evaluate the performance of our thesaurus-building methods. By establishing a solid foundation of gold-standard synonyms, we could measure the accuracy and usefulness of our thesaurus, and identify areas for improvement as we refined our approach.

4 Experiment 1: Welsh word embeddings

Word embeddings are widely used in NLP and machine learning tasks to capture the semantic and syntactic properties of words in a continuous vector space. FastText is a popular method for training word embeddings that can handle out-of-vocabulary words and subword information using character n-grams (Grave et al., 2018). The two pre-trained Welsh word embeddings used in this experiment were the FastText embeddings trained on a large Welsh corpus from Wikipedia and the fine-tuned FastText (Fine-Tuned-FastText) embeddings using the Welsh Wikipedia as well as the Welsh National Corpus along with 9 other resources (92,963,671 words) (Corcoran et al., 2021).

To evaluate the performance of the word embeddings, we used the gold-standard synonyms generated by Welsh speakers as the reference. We compared the generated synonyms for each word in the gold-standard with the synonyms generated by the two word embeddings.

We used the FastText embeddings and fine-tuned-FastText to generate the 10 nearest (most related) words to each input word on our 250-word list. The resulting list of nearest words for the example word “pobl” is shown in Table 1, along with their translations. Based on the Table 1, it is clear that the fine-tuned FastText approach yielded better results than the standard FastText approach in terms of identifying the most related words to the Welsh word “pobl”. The most related words generated by the fine-tuned FastText approach were very close in meaning to the original word, as indicated by their high similarity scores ranging from 0.733 to 0.468. In contrast, the most related words generated by the standard FastText approach had lower similarity scores ranging from 0.629 to 0.504.

An important point to consider is that the nearest words may include antonyms of the input word, as the embeddings are based on the behaviour of the word in various contexts. This process allowed us to leverage the power of FastText embeddings to quickly and automatically generate potential synonyms for each word on our list.

To refine the word embedding results, we used the Python Multilingual UCREL Semantic Analysis System (PyMUSAS)[7], which retains Welsh language resources and methods originally included in an earlier Java version developed during the CorCenCC project (Piao et al., 2018). The PyMUSAS tagger assigns a set of fine-grained semantic tags to each word based on its POS (assigned by the CyTag Welsh POS tagger also created during the CorCenCC project), morphological features, and semantic field. We selected a subset of the generated fastText words for each original word based on matching the semantic tags and removing matching lemmas. This can be done by comparing the

[7] https://pypi.org/project/pymusas/

FastText			Fine-Tuned-FastText		
Vector	**Most Rel.**	**Trans.**	**Vector**	**Most Rel.**	**Trans.**
0.629	bobl	people	0.733	pobol	people
0.556	rhai	some	0.641	bobl	people
0.546	phobl	people	0.554	phobl	people
0.530	LHDTQ	LGBTQ	0.551	bobol	people
0.528	pobol	people	0.551	rhywun	someone
0.522	cleiantiaid	clients	0.540	trigolion	inhabitants
0.515	ifanc	young	0.498	pawb	everyone
0.515	bod	being	0.482	dinasyddion	citizens
0.514	trwy'r	through/by the	0.480	plant	children
0.504	Ogleddwyr	North Walians	0.468	pobl'	people

Table 1: 10 most related words to the Welsh word 'pobl'

semantic tags of the generated words with the semantic tag(s) of the original word and selecting the ones that share the same tag(s). Table 2 compares the performance of FastText embeddings and Fine-Tuned-FastText word embeddings in finding synonyms for the Welsh word "pobl". While not all the FastText embeddings share the POS tag and seven synonyms do not share the PyMUSAS tag with the original word. Seven synonyms share both the POS tag and PyMUSAS tag of the original word, and three of the new synonyms share the same lemma with the original word. Based on this analysis, it appears that the Fine-Tuned-FastText word embeddings perform better than the FastText embeddings in terms of producing synonyms that share the same POS speech tag and PyMUSAS tag as the original word. Therefore, it may be worth exploring different techniques for refining word embeddings' results, such as using a lemmatiser and semantic tagging.

In this experiment, when the semantic tagger produced a $Z99$ for a word that was not in its lexicon, the approach taken was to remove only the matched lemmas from the list rather than eliminating the $Z99$ words to avoid a very short list.

After applying lemmatisation and removing words that share the same lemma as the original word, the number of data entries was reduced from 2490 to 2047 for the FastText model, with an average of fewer than 9 synonyms per word. For the fine-tuned FastText model, the number was reduced to 1776, with an average of 7 synonyms per word. The lemmas that exactly match the original word lemma are in-bold font in Table 2.

Next, by selecting only the words that share the same PyMUSAS semantic tag but do not share the lemma, the number of entries further reduced to 132 for the FastText model and 173 for the Fine-Tuned-FastText model. This means that some of the data did not have any synonyms that share the same semantic PyMUSAS tag [S2, People].

This process of selecting synonyms that share the same semantic tag but not the same lemma can be useful in reducing redundancy and increasing the diversity of the synonyms list. It can also help in avoiding circular dependencies and improving the quality of the generated data. However, it is important to note that this process may also result in a loss of some relevant synonyms that do not share the same semantic tag as the original word. Therefore, it is essential to carefully evaluate the trade-offs and choose the appropriate method.

Once we have selected a subset of the generated words based on semantic similarity and lemma dissimilarity, we can match them with the gold-standard user input by comparing the words and their order with the user-generated synonyms. This will allow us to evaluate the quality and relevance of the generated synonyms and identify any discrepancies or inconsistencies with the user input.

5 Experiment 2: Analysis to create gold-standard

The objective of this study was to analyse the input provided by Welsh speakers in generating synonyms for a pre-compiled list of 250 Welsh words. The study aimed to create a gold-standard list of synonyms for these words based on the input of seven paid evaluators for each word. The seven evaluators were native speakers of Welsh, either in the final year of an undergraduate Welsh degree programme or postgraduate students with experi-

FastText				Fine-Tuned-FastText			
Word	**Lemma**	**POS**	**PyMUSAS**	**Word**	**lemma**	**POS**	**PyMUSAS**
bobl	**pobl**	E	S2	pobol	pobol	E	**S2**
rhai	rhai	unk	A13.5	**bobl**	**pobl**	E	S2
phobl	**pobl**	E	S2	**phobl**	**pobl**	E	S2
LHDTQ	LHDTQ	E	Z99	bobol	pobol	E	**S2**
pobol	pobol	E	**S2**	rhywun	rhywun	E	Z8mfc
cleiantiaid	cleiantiaid	unk	Z99	trigolion	trigolyn	E	H4/S2mf
ifanc	ifanc	Ans	T3-	pawb	pawb	unk	Z8/N5.1+c
bod	bod	B	A3+, Z5	dinasyddion	dinesydd	E	G1.1/S2mf
trwyr	trwyr	unk	Z99	plant	plentyn	E	S2mf/T3
Ogleddwyr	Ogleddwyr	E	Z99	**pobl**	**pobl**	E	S2

Table 2: Comparison of FastText and Fine-Tuned-FastText Word Embeddings in Generating Synonyms for the Welsh Word 'pobl'

ence of writing in Welsh. They were asked to provide up to ten synonyms for each word, and the order in which they presented the synonyms was determined individually.

To create the gold-standard list, the study conducted comparison experiments to match the agreement of synonyms and their POS across the evaluators, as well as the agreement of the ordering of the presented synonyms. Additionally, the ordering of the synonyms provided by the evaluators was compared against the frequency of these words in the CorCenCC corpus frequency.

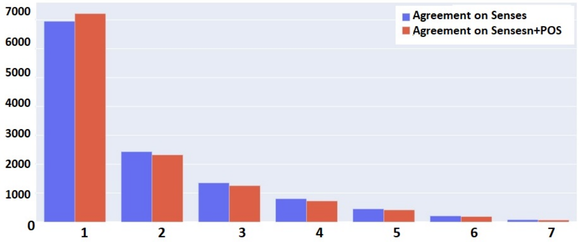


Figure 1: Sense versus POS agreement across the participants

Table 3 presents information on the level of agreement among annotators on the senses of words, with and without their part of speech tags. The table shows that for the majority of the words (4517 out of 6953), only one annotator suggested the sense of the word. As the number of annotators in agreement increases, the number of words decreases, indicating that agreement among annotators is less common for most of the words. For instance, only 64 words had seven annotators in agreement on their sense and part of speech tag. Figure 1 represents the relationship between sense and POS agreement across participants.

The overall aim of the study was to provide a reliable and standardised list of synonyms for Welsh words that could be used in NLP applications. By analysing the input of multiple evaluators and creating a gold-standard list based on their input, the study aimed to ensure that the list was comprehensive and accurate. Furthermore, the study aimed to provide insights into the agreement and variability among Welsh speakers in generating synonyms, as well as the relationship between the ordering of synonyms and their frequency in the corpus. This experiment can be used to further develop NLP applications for Welsh language processing tasks and improve the accuracy and relevance of the generated synonyms.

Agreements	Sense	Sense & POS
only 1	4517	4895
at least 2	2436	2323
at least 3	1354	1257
at least 4	808	729
at least 5	454	416
at least 6	209	186
at least 7	80	64

Table 3: Gold-standard Words Agreements.
Agreements: the number of annotators in agreement. Sense: number of agreements on senses. Sense & POS: number of agreements on senses and their part of speech tags

The gold-standard synonyms provided by the seven participants were ordered based on the mean position of each synonym across all participants. For instance, the word "pobl" had 31 unique synonyms suggested by the participants, as shown in Figure 2. To quantify the variability or fluctua-

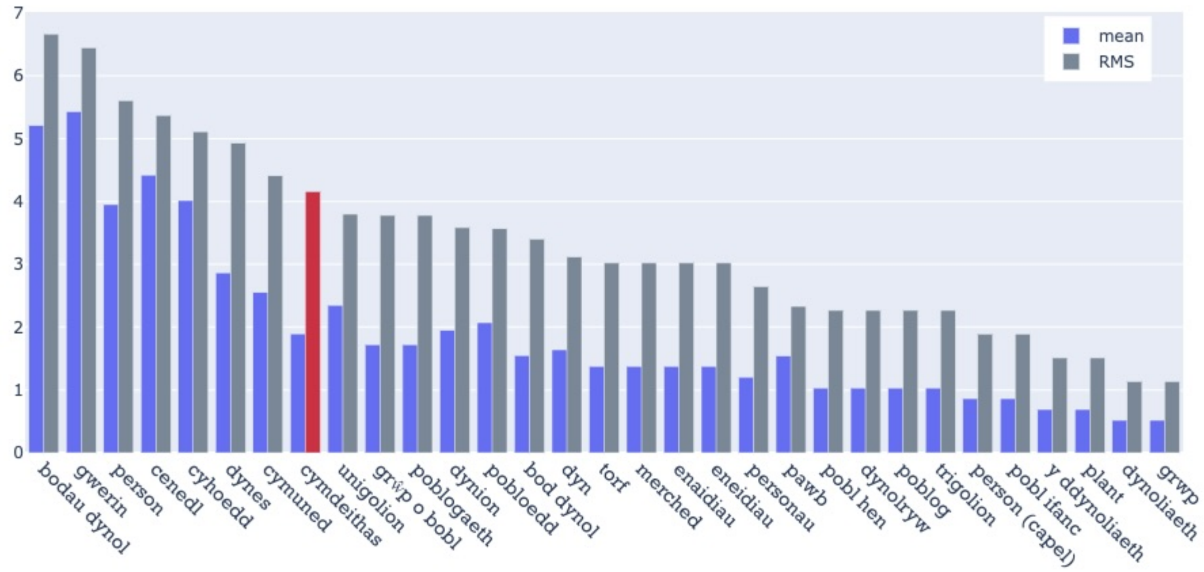


Figure 2: Mean versus RMS for the word 'pobl' senses

tion in a set of values, we used the Root Mean Square (RMS), which is a mathematical measure commonly used in various fields, including language processing. Let a set of n values be denoted by

$$x_1, x_2, ...x_n$$

Then, the RMS can be computed as:

$$x_{RMS} = \sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + x_3^2.... + x_n^2\right)}$$

By using RMS to reorder synonyms, words suggested by a single participant but in a higher position will be given more weight than words suggested by multiple participants but in lower positions. This is because RMS takes into account the variability of the data and gives more weight to values that are farther from the mean.

In the specific example of the word "cymdeithas" [red column] shown in Figure 2, the word was introduced by only one speaker but was in a higher position when RMS was used to reorder the synonyms. This indicates that the word was used more frequently or prominently by the participant who suggested it, and thus should be given more weight in the final output.

Using RMS to reorder synonyms can be a useful technique to ensure that the most relevant and frequently used words are given priority, even if they are suggested by fewer participants. This approach can help to produce a more accurate and representative list of synonyms for Welsh words, which can be valuable for various NLP applications.

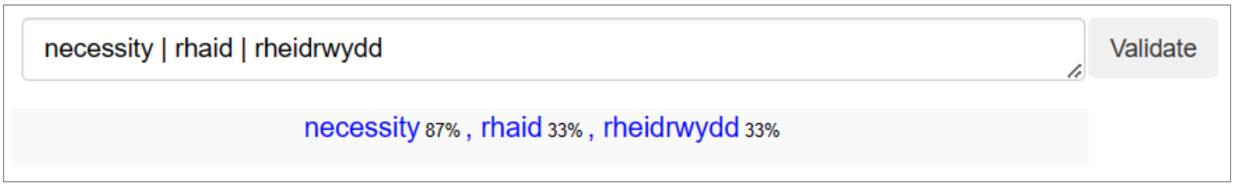
In this case the 250 list of words was expanded by adding 6953 synonyms from linguists, and we now have a more extensive words to build the gold-standards. With this expanded list of words, when a user queries a particular word, the thesaurus can now present all of its synonyms as well, allowing the search to see and choose from a wider range of options. This can be especially useful when a user is unsure of the exact word they want to use, or when they want to explore different ways to express a particular concept.

One thing to keep in mind is that not all synonyms are interchangeable in every context, and some synonyms may have different connotations. Therefore, it was important to consider the context in which each synonym is used and to provide additional information or context as needed to help users choose the most appropriate synonym for their particular situation. This will be done by extracting an example for each word from CorCenCC corpus (Knight, 2020; Knight et al., 2021).

6 Experiment 3: Graph-based Approach

For our next experiment utilising existing dictionaries and thesauri, we developed a web tool for validating Welsh synonyms based on a graph-based algorithm as described by Ghanem et al. (2023)[8]. This algorithm constructs a graph at level *k* from a set of translation or synonymy pairs and consid-

[8] https://portal.sina.birzeit.edu/synonyms/



necessity | rhaid | rheidrwydd

necessity 87% , rhaid 33% , rheidrwydd 33%

Figure 3: Synonyms Web Tool

ers all cyclic paths as candidate synonyms. The algorithm then calculates a fuzzy value for each candidate synonym to determine its likelihood of being a member of a synset.

Figure 3 depicts the tool[9], which features several bilingual dictionaries, including the Welsh-English Dictionary by Hawke and the Welsh WordNet[10], that we uploaded to the tool. It accepts a set of synonyms and validates them using this algorithm.

Table 4 displays the assessment of the linguists' synonyms in comparison to the tool's outcomes using three evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Jaccard coefficient.

MAE and RMSE are numerical prediction accuracy metrics used to determine the ranking difference between the linguists' synonyms and the tool's results, with the synonyms represented by numerical vectors. The MAE measures the average absolute difference between the predicted and actual values, while the RMSE measures the square root of the average squared differences between the predicted and actual values. The evaluation resulted in MAE values ranging from 10.02 to 28.79 and RMSE values ranging from 13.05 to 35.38. Linguist 4 at level 3 exhibited the lowest MAE and RMSE values of 19.26 and 23.36, respectively, signifying the highest level of synonym prediction accuracy. Overall, the performance of all linguists and word-embeddings (WE) was superior at level 3 than at level 2.

The Jaccard coefficient calculates the similarity between two sets, ranging from 1 for identical sets to 0 for completely dissimilar sets. If a synonym is not found in the tool, it is ranked at the end of the synset and labeled as "out of vocabulary. Consequently, we must measure the overlap between the tool's identified synonyms and the input synset using the Jaccard coefficient. The comparison outcomes varied from 0.34 to 0.83, with linguist 4 at level 3 exhibiting the highest Jaccard coefficient value of 0.83, indicating a high degree of similarity between their synonyms and the reference set.

Linguist	Level	Jaccard	MAE	RMSE
1	2	0.77	28.55	35.00
	3	0.80	28.79	35.38
2	2	0.75	22.34	27.18
	3	0.77	22.34	27.40
3	2	0.74	24.27	30.01
	3	0.76	23.89	29.74
4	2	0.81	19.11	23.15
	3	**0.83**	19.26	23.36
5	2	0.75	24.06	29.53
	3	0.77	23.83	29.39
6	2	0.52	14.54	18.15
	3	0.54	14.86	18.51
7	2	0.34	10.02	13.05
	3	0.35	10.03	13.07
WE	2	0.43	27.69	33.79
	3	0.45	28.74	35.03

Table 4: Evaluation of linguists' synonyms against the dictionaries and WordNet results

Overall, the evaluation results indicate significant variation in the quality of the linguists' synonyms, with linguist 4 at level 3 demonstrating the best performance across all three evaluation metrics.

This experiment provides valuable insights into the effectiveness of multilingual extraction methods in generating related words in Welsh, while also highlighting the strengths and limitations of different techniques and linguists. These findings can further inform the development of more precise and comprehensive thesauri and word embeddings for Welsh language processing tasks.

7 Conclusion and Future Work

In this paper, we presented our approach to creating a comprehensive thesaurus for Welsh using a combination of existing resources and novel techniques. We demonstrated the effectiveness of our

[9] https://portal.sina.birzeit.edu/synonyms/

[10] https://datainnovation.cardiff.ac.uk/is/wecy/access.html

approach through a series of experiments and evaluations, and showed that our thesaurus outperformed existing Welsh-language resources in generating related words. Our approach leverages the power of FastText embeddings and semantic tagging to generate candidate synonyms, and RMS reordering to identify the most relevant and frequently used words.

However, our work is not without limitations. While we aimed to create a comprehensive and accurate thesaurus, it is possible that our resource is still incomplete or may contain errors. To further refine and improve the thesaurus in the future, we will enlist the help of human evaluators who are fluent in Welsh. Specifically, we will use a pre-existing platform to crowd-source human participants to evaluate the resource. This will help ensure that the thesaurus is relevant, accurate, and meets the needs of its users, enhancing its value and utility for Welsh speakers and learners.

By combining or comparing the results of the three experiments, we can gain a deeper understanding of how to optimise our approach and further refine the thesaurus. Specifically, we can identify areas for improvement and investigate how to address potential limitations or errors in the resource.

Overall, our work contributes to the growing body of research on NLP and machine learning for under-resourced languages, and demonstrates the potential of using novel techniques and approaches to create valuable resources for these languages. We hope that our work will inspire further research and development in this area, and that our thesaurus will be a useful tool for Welsh speakers, learners, and researchers alike.

Acknowledgements

We would like to express our gratitude to the evaluators who participated in this study, all of whom are native Welsh speakers. Their invaluable contributions and dedication to the project made this research possible. This research was funded by the Welsh Government, under the Grant 'Using Word Embeddings to Create an Interactive Thesaurus of Contemporary Welsh'.

8 Ethics

The payment provided to the evaluators was in accordance with the UK's national minimum wage regulations, to ensure that it meets or exceeds the standard wage requirements.

References

Jean Aitchison, Alan Gilchrist, and David Bawden. 2000. *Thesaurus Construction and Use: A Practical Manual*, 4th edition. Aslib, London.

Rudy Arthur and Hywel TP Williams. 2019. The human geography of twitter: Quantifying regional identity and inter-region communication in england and wales. *PloS one*, 14(4):e0214466.

Patrick Carlin and Diarmait Mac Giolla Chríost. 2016. A standard for language? policy, territory, and constitutionality in a devolving wales. In *Sociolinguistics in Wales*, pages 93–119. Springer.

Padraig Corcoran, Geraint Palmer, Laura Arman, Dawn Knight, and Irena Spasić. 2021. Creating welsh language word embeddings. *Applied Sciences*, 11(15):6896.

Daniel Cunliffe, Delyth Morris, and Cynog Prys. 2013. Young bilinguals' language behaviour in social networking sites: The use of welsh on facebook. *Journal of Computer-Mediated Communication*, 18(3):339–361.

Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris, and Dawn Knight. 2022a. Creation of an evaluation corpus and baseline evaluation scores for welsh text summarisation. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 14–21.

Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris, and Dawn Knight. 2022b. Welsh summaries correlation between rouge and human evaluation. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 14–21.

Ignatius Ezeani, Mahmoud El-Haj, Jonathan Morris, and Dawn Knight. 2022. Introducing the Welsh text summarisation dataset and baseline systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5097–5106, Marseille, France. European Language Resources Association.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT press.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. *The 12th International Global Wordnet Conference (GWC2023).*

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Amir Hazem and Béatrice Daille. 2018. Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dawn Knight. 2020. Corcencc: Corpws cenedlaethol cymraeg cyfoes–the national corpus of contemporary welsh. *Oxford Text Archive Core Collection*.

Dawn Knight, Fernando Loizides, Steven Neale, Laurence Anthony, and Irena Spasić. 2021. Developing computational infrastructure for the corcencc corpus: the national corpus of contemporary welsh. *Language Resources and Evaluation*, 55:789–816.

Dawn Knight, Steve Morris, Beth Tovey-Walsh, and Tess Fitzpatrick. 2020. Yr amliadur: Frequency lists for contemporary welsh.

Jörg Landthaler, Bernhard Waltl, Dominik Huth, Daniel Braun, and Florian Matthes. 2018. Extending thesauri using word embeddings and the intersection method. In *Proceedings of the 10th International Conference on Knowledge Engineering and Ontology Development*, pages 159–166. SciTePress.

Bernardo Magnini, Carlo Strapparava, Piotr Pezik, and Ido Dagan. 2000. Integrating subject field codes in wordnet. *Computational Linguistics*, 26(2):199–227.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Jonathan Morris, Ignatius Ezeani, Ianto Gruffydd, Katharine Young, Lynne Davies, Mahmoud El-Haj, and Dawn Knight. 2022. Welsh automatic text summarisation. In *Wales Academic Symposium on Language Technologies*. Banolfan Bedwyr.

Scott Piao, Paul Rayson, Dawn Knight, and Gareth Watkins. 2018. Towards a Welsh semantic annotation system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Scott Piao, Paul Rayson, Dawn Knight, Gareth Watkins, and Kevin Donnelly. 2017. Towards a welsh semantic tagger: creating lexicons for a resource poor language. In *Proceedings of Corpus Linguistics*.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL Semantic Analysis System. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 7–12.

Peter D Turney. 2007. Measuring semantic similarity by latent relational analysis. *The Journal of Computer and System Sciences*, 73(4):389–401.

Cen Williams. 1999. *Cymraeg Clir: Canllawiau Iaith*. Bangor: Gwynedd Council, Welsh Language Board and Canolfan Bedwyr.