A Major Wordnet for a Minority Language: Scottish Gaelic

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2812–2818
Marseille, 11–16 May 2020

A Major Wordnet for a Minority Language: Scottish Gaelic

Gábor Bella[1], Fiona McNeill[2], Rody Gorman[3], Caoimhín Ó Donnaíle[3], Kirsty MacDonald, Yamini Chandrashekar[1], Abed Alhakim Freihat[1], and Fausto Giunchiglia[1]
[1]University of Trento, via Sommarive, 5, 38123 Trento, Italy
[2]Heriot–Watt University, Edinburgh, EH14 4AS, Scotland
[3]Sabhal Mòr Ostaig, University of the Highlands and Islands, Sleat, Isle of Skye, IV44 8RQ, Scotland
gabor.bella@unitn.it, f.mcneill@hw.ac.uk, anguth@btinternet.com, caoimhin@smo.uhi.ac.uk, curstag@gmail.com, yamini.chandrashekar@studenti.unitn.it, abdel.fraihat@gmail.com, fausto.giunchiglia@unitn.it

Abstract

We present a new wordnet resource for Scottish Gaelic, a Celtic minority language spoken by about 60,000 speakers, most of whom live in Northwestern Scotland. The wordnet contains over 15 thousand word senses and was constructed by merging ten thousand new, high-quality translations, provided and validated by language experts, with an existing wordnet derived from Wiktionary. This new, considerably extended wordnet—currently among the 30 largest in the world—targets multiple communities: language speakers and learners; linguists; computer scientists solving problems related to natural language processing. By publishing it as a freely downloadable resource, we hope to contribute to the long-term preservation of Scottish Gaelic as a living language, both offline and on the Web.

Keywords: wordnet, Scottish Gaelic, under-resourced language, minority language, lexical semantics, translation, lexical gap, language diversity

1. Introduction

Scottish Gaelic, a Celtic language, derives from Middle Irish, yet it is considered today as a distinct language. While it used to be spoken throughout Scotland, its current speakers are estimated to be fewer than 60,000 (Census, 2011) and is considered as an endangered language.

As the World Wide Web takes over ever higher portions of our everyday life, the accessibility, and ultimately the survival, of pretty much any idea or artifact is increasingly determined by its online visibility. As the seminal article on *digital language death* points out (Kornai, 2013), this observation also holds for languages: those that are not actively used online will likely fall behind in their deemed usefulness by its potential speakers, accelerating their extinction.

For most minority languages, a further problem arises from the often incomplete lexicon with respect to modern terms, pertaining to inventions, discoveries, and other phenomena from the 20th and 21st centuries. The lack of suitable terms to designate contemporary concepts is perceived by potential speakers as a hindrance to effective language use.

Similar problems are faced from a computational perspective: the weak online presence of minority languages, such as Scottish Gaelic (*Gaelic* in short in the rest of the paper), transpires as a lack of digital corpora, which prevents state-of-the-art data-driven language processing methods, such as machine translation, from being applied in an efficient manner.

As an attempt to counter these tendencies, we release the *Unified Scottish Gaelic Wordnet*, a free lexico-semantic resource containing over 10k words that lexicalise 13k word meanings (*synsets* in wordnet terminology).[1] The resource was built by merging two sources: in a larger part (about 60%), the translation of subsets of the English *Princeton WordNet*[2] (Miller, 1998) by language experts; and in a smaller part (about 40%), an existing wordnet from the *Extended Open Multilingual Wordnet* project (Bond and Foster, 2013a), itself directly converted from the Scottish Gaelic Wiktionary.

Contrary to wordnets in most other languages, our wordnet also provides over 600 explicitly marked Gaelic *lexical gaps* (English words that have no Gaelic equivalent) as well as 73 English gaps (Gaelic words that have no English equivalent). This—for the moment relatively small yet unique—set of gaps has good potential to be exploited in further research on language diversity (Giunchiglia et al., 2017).

For automating the generation of translation tasks (to be undertaken by human language experts), computing statistics, merging the two wordnets, and generating the end result, we used as general framework the *Universal Knowledge Core* (UKC), a large-scale multilingual lexico-semantic resource that currently consists of the lexicons of over a thousand languages, represented as wordnet structures.[3]

We consider this wordnet to be potentially beneficial for the continued use of the Gaelic language. Due to being linked to all other wordnets of the world, it can be exploited by humans as a multilingual dictionary. It can also be used computationally, e.g. for natural language understanding tasks (such as word sense disambiguation on Gaelic text) or as a seed dictionary for the generation of cross-lingual word embeddings. The latter may, in turn, be able mitigate the lack of large Gaelic corpora for learning-based solutions.

The rest of the paper is organised as follows: section 2 presents the state of the art on lexico-semantic resources for Gaelic. Section 3 provides a quick overview on wordnets and on how the UKC was used to drive wordnet genera-

[1] http://ukc.disi.unitn.it/index.php/gaelic/

[2] Following conventional usage, we use the orthography *'WordNet'* to refer to the original English Princeton WordNet, and *'wordnet'* to designate similar resources of other languages.

[3] http://ukc.disi.unitn.it/

tion. Section 4 gives details on the methodology used for producing the wordnet. Section 5 presents the end results, including statistics, availability, and the lessons learnt. Finally, section 6 reflects on follow-up work.

2. Lexical Resources for Gaelic

Although Celtic languages dominated Europe in the late centuries BCE up until the spread of Latin throughout the continent, there are currently only six extant languages[4]. These are divided into the Brittonic or P–Celtic languages—Welsh, Breton, and Cornish—and the Goidelic or Q–Celtic languages—Scottish Gaelic, Irish,[5] and Manx. The latter three all derive from Middle Irish. Whilst many similarities still exist between the Goidelic languages, they are not generally mutually comprehensible[6], and there are significant lexical, phonetic, grammatical, and orthographic differences. The Brittonic and Goidelic languages are significantly different and not mutually comprehensible.

2.1. State of the Art on Gaelic Resources

There is a shortage of resources for all of these minority languages. Irish, Welsh and Gaelic are all recognised as official languages in their indigenous countries (Ireland, Wales and Scotland respectively) and receive some level of state support, including provision of state education through the medium of the language from 3–18 and in higher education[7]. Nevertheless, developing extensive resources for minority languages is challenging.

Though Gaelic has been a written language for many hundreds of years, the first attempt to create a significant Gaelic–English dictionary was the *Armstrong dictionary* of 1825 (Armstrong, 1825), which was quickly followed by the larger *Dictionarium Scoto-Celticum* (Maclachlan, 1828), created in 1828 by the Highland Society of London, which provided translations of Gaelic words into both English and Latin. In 1901, Edward Dwelly produced the *Dwelly dictionary* (Dwelley, 1990), which contains over 70,000 entries and is still widely considered to be the most comprehensive dictionary of the language compiled to date (McLeod, 2013). Most existing online resources are based to some extent on the Dwelly dictionary, such as its digitised version *Dwelly-d*.[8] Probably the most widely used online dictionary at present is *Am Faclair Beag*[9], built from Dwelly and other sources by Michael Bauer. It was later adapted to create the *Learn Gaelic* online dictionary[10], which is particularly aimed at learners and is the main language resource used in Gaelic-medium schools.

There are two different wiktionaries for Gaelic. Under `gd.wiktionary.org` is a Gaelic-medium dictionary with definitions of Gaelic words in Gaelic (and some other languages). It only contains around 188 Gaelic words. The wiktionary under `en.wiktionary.org` has definitions of Gaelic words given in English. It has recently been improved and contains 8,638 entries. Despite being of decent quality, it is not widely used, primarily because it is less complete than other Gaelic dictionaries.

Glosbe is a multilingual online dictionary.[11] Currently, its English–Gaelic dictionary contains about 12,904 translated phrases, and 1,037 translated sentences, aggregated from various sources: mostly wiktionaries, but also machine translations and crowsourced translations. Contrary to conventional dictionaries, this resource also contains many proper nouns and longer phrases.

In addition to generic dictionaries, there are multiple linguistic resources with a particular specialism. *An Stòr-dàta Briathrachais* is a database of around 100,000 word pairs primarily focussing on technical terms, but also containing many general terms, that was developed at *Sabhal Mòr Ostaig* in the early 1990s and could be considered the first semantic resource in Gaelic. A thesaurus was recently created for the Learn Gaelic dictionary,[12] commissioned by the Scottish Government and produced by hand by experts. There are also dictionaries available that focus on the natural world, and *Faclair Riaghaltas Ionadail* contains government terminology. *An Sruth*[13] focusses on phrases and idioms. It was originally developed as a Gaelic–Irish resource, primarily by the fourth author, but English translations have also been added.

All of these resources and several others can be accessed via the Multidict interface[14] (Ó Donnaíle, 2014), allowing users to find translations via many different resources—though in practice, only the major resources are frequently used.

Finally, the *Extended Open Multilingual WordNet* project (Bond and Foster, 2013b) was the first to create a Gaelic wordnet, along with over 150 other languages. This was done automatically, by extracting data from the Gaelic Wiktionary and aligning it with Princeton WordNet synsets. This created a wordnet with 5,498 synsets and 4,674 words; however, Gaelic-language glosses are completely absent from this resource.

In comparison to these efforts, our goal was to provide a lexico-semantic resource that is of a usable size to cover a considerable part of the common vocabulary (even if far from being exhaustive in its initial stage), that is sense-aligned not only with English but also with other languages, that is of high quality due to human supervision, and finally that is exploitable both computationally and by humans. We chose the wordnet format, commonly used for computational tasks (Agirre and Edmonds, 2007; Bella et al., 2016), to satisfy this last criterion.

[4] Arguably only four: both Cornish and Manx died in modern times but revitalisation efforts have led to existing native speakers.

[5] The Irish name for the language is *Gaeilge* and it is sometimes referred to as *Irish Gaelic* or even just *Gaelic* in English (hence the need to specify Scottish Gaelic), but *Irish* is the preferred term. Likewise Manx is sometimes referred to as Manx Gaelic

[6] This differs to some extent between different dialects, with Scottish Gaelic closer to Ulster Irish than Irish from more southerly or westerly districts

[7] In Scotland, approximately 1.6% of school-aged children (11,103) were in Gaelic-medium education in 2018.

[8] `http://www.dwelly.info`

[9] `https://www.faclair.com`

[10] `https://learngaelic.scot/dictionary/`

[11] `https://glosbe.com`

[12] `https://www.learngaelic.net/thesaurus/`

[13] `http://www.smo.uhi.ac.uk/teanga/sruth/`

[14] `https://multidict.net/multidict/`

2.2. Wordnet Creation for Minority Languages

(Vossen, 1998) argues that there are two major, fundamentally different ways of creating new wordnets, through what he calls the *expand* and the *merge* approaches. The first takes an existing wordnet as basis—usually the English Princeton WordNet as it is the most complete—and proceeds by providing translations for a carefully selected subset of synsets, based on both the source words and the gloss. Examples of efforts using this approach are (Pociello et al., 2011) for Basque or (Ganbold et al., 2018) for Mongolian. The expand approach may be implemented in different modalities, such as expert sourcing (as in our case) or crowdsourcing as in (Ganbold et al., 2018).

The second major approach takes one or more existing resources instead, such as monolingual thesauri and/or bilingual dictionaries, and builds a new synset hierarchy that is usually different from that of Princeton WordNet. This is the approach used by the wordnets generated in the *Open Multilingual Wordnet* project (Bond and Foster, 2013b).

The choice of approach has a fundamental effect on the end result: in the case of expansion (translation), the new wordnet will be fully meaning-aligned with the source language (English), which is ideal for cross-lingual uses: as most wordnets are already aligned with PWN, we get bilingual translations to all those languages 'for free'. On the other hand, a certain linguistic bias is introduced by the fact that only meanings for which English lexicalisations exist will appear in the wordnet. In other terms, words culturally specific to Gaelic are likely to be omitted from the end result. In contrast, the merge approach may produce a less biased representation of the language; however, it is much harder to map *a posteriori* in a precise way to other languages.

Our work has adopted the expand approach as we considered interoperability with other languages a priority from a point of view of language preservation and the overall usefulness of the resource. A precise mapping to English was also necessary in order to be able to merge our translations with Wiktionary. Nevertheless, as we will show below, we did also address certain aspects of language diversity by explicitly representing *lexical gaps*: English words that have no lexicalisation in Gaelic and, vice versa, words that are specific to the Gaelic language and culture without English equivalents.

3. Wordnets and the UKC

This section provides a brief background on the wordnet data structure, as defined for the original Princeton WordNet (PWN) by (Miller, 1998) and used more or less identically for hundreds of other languages. We also provide an overview of the UKC framework that we used to automate various steps of our work. For more details, we refer the reader to the respective articles (Miller, 1998) and (Giunchiglia et al., 2018).

Wordnets are rich and complex graphs that represent the lexicon of a language—the *words*—as well as word meanings formalised as *synsets*—sets of synonyms. Synsets are organised into hierarchies according to lexico-semantic relations such as hypernymy, meronymy, and troponymy. Wordnets often also provide other kinds of relations or classifications of lexical items.

The Universal Knowledge Core aggregates and extends the wordnets of the world (Giunchiglia et al., 2018). It currently contains the wordnets of 340 languages. Beyond merely being a wordnet aggregator, the UKC provides importing, merging, and exporting mechanisms for language resources that we exploited in our work as described in the section below. It also extends the monolingual wordnet structure by cross-lingual knowledge (Batsuren et al., 2019), including a *concept layer* that reifies cross-lingual equivalence relations among synsets (word meanings) into supra-lingual concepts. It thus provides an effective word translation mechanism across all of its languages, which we exploit for the production of the Unified Scottish Gaelic Wordnet.

4. Methodology

As evoked in section 2, we have adopted the expert-sourced expand approach to building our wordnet resource, i.e. a subset of words, glosses, and examples from the English Princeton WordNet were translated and validated by Gaelic language experts. Accordingly, the macro-steps of our methodology were as follows:

1. *translation task generation:* first we specified which subset of PWN to translate;
2. *translation:* the actual translation effort carried out by a Gaelic language expert;
3. *merge:* fusing the translation results with the words provided by the existing Wiktionary-based Gaelic wordnet;
4. *validation:* a subset of the translated terms was evaluated and corrected by a different language expert.

4.1. Translation Task Generation

Wordnets are directed acyclic graphs where nodes correspond to synsets and edges to hyponymy relations. The graph is not entirely connected: the four parts of speech covered—nouns, verbs, adjectives, and adverbs—form separate partitions. In PWN 2.1, the proportion of noun, verb, adjective, and adverb senses is, respectively, 70%, 12%, 15%, and 3%: nouns clearly are the bulk of it. The noun graph itself is a connected graph with a single root node meaning 'entity' and a maximum depth of less than 20. Nodes (synsets) closest to the root are often abstract philosophical concepts such as 'physical object' or 'stative'. Right below one typically finds word meanings that, according to Rosch's cognitive theory (Rosch, 1999), can be qualified as *basic level categories*: common everyday concepts such as 'dog', 'house', etc. Towards the bottom of the graph one enters into domain territory with a large number of specialised terms from medicine, zoology, etc.

Beyond the obvious constraint of the translator's limited availability, the following criteria were used when selecting the subset of English synsets to be translated:

- favour general language as opposed to domain terms;
- yet, cover much of the grey area between basic-level categories and domain terms, in order to increase the

vocabulary coverage of Gaelic with potential neologisms;

- do not overlap significantly with the Wiktionary-based wordnet (small overlaps were kept for cross-validation purposes);
- favour nouns and verbs (adjectives and adverbs would be translated in a later phase).

Translation tasks (i.e. subsets of synsets) were defined in terms of *subtrees*, that is, a root node that corresponds to a very general category together with all of its descendants. 13 such subtrees of noun synsets were selected, underneath categories such as *natural object*, *body part*, *feeling*, *event*, *food*, *location*, etc. Finally, a 14th set contained about 1,100 verb synsets corresponding to more commonly used verbs, as decided by language expert judgment.
The translation tasks were generated by exporting language data from the UKC as spreadsheets (one per subtree). One spreadsheet row was generated for each synset, containing the English synset ID, the English source lemmas, gloss, and example phrases, as well as empty slots for introducing the translated lemmas, glosses, and examples. The synsets were output in breadth-first order in order for the task to proceed from the more general gradually towards the more specific meanings.

4.2. Translation

Translations were provided by the third author, a Gaelic professional translator (and poet and writer) with a strong past experience in dictionary translation for the Scottish Government. The following are the most notable instructions he was given:

- he was given the authority to decide to skip synsets or to stop translating a subtree entirely when he deemed the terms were becoming too technical and consequently of limited interest to non-specialist users;
- he was asked explicitly to identify *lexical gaps* (where no Gaelic lexicalisation exists) and take one of the following actions:
  - mark the synset as a lexical gap, while still providing a Gaelic gloss for it (that provides an approximation or explanation of the meaning),
  - invent a neologism, a new Gaelic word, clearly marking it for later identifiability.

Examples of Gaelic lexical gaps are *'spoonerism'* (i.e. the transposition of initial consonants in a pair of words), or *'to launder'* (money). Marking such non-existent Gaelic words as lexical gaps is a major feature of the wordnet, for several reasons. First of all, a clear distinction is made with respect to resource incompleteness: such Gaelic translations are not merely missing from the wordnet, but from the Gaelic lexicon itself. Secondly, lexical gaps are prime examples of cross-lingual *lexical diversity* and as such are a good starting point for diversity-related linguistic studies (Giunchiglia et al., 2017).

The overall translation effort resulted in 10,583 word senses (words with one specific meaning) considered by the translator, out of which 1,614 were either skipped because they were too specialised or, in 733 cases, were marked up as lexical gaps in Gaelic. The remaining 8,969 senses were translated either into one or more existing Gaelic words (6,576, 74% of the translations) or else covered by a neologism coined by the translator (2,393, 26%). Furthermore, 8,030 synsets (97.4%) were provided a Gaelic gloss, and 264 (3.2%) of them an example phrase (the latter number is low with respect to Princeton WordNet, but is similar to other wordnets of the world that typically lack example phrases).

Most neologisms were created in a conservative manner: either as direct literal translations of the English lemmas, or using derivation. For example, new verbs were derived from nouns by adding the Gaelic word *déan* (to do / to make something) as prefix or the suffix *-ich* (similar to the English -ise / -ize) to existing words.

4.3. Merge

For the purpose of merging the translations with the Wiktionary-based Gaelic wordnet—produced by the *Extended Open Multilingual Wordnet* (EOMW) project (Bond and Foster, 2013a)—we used the importing and merge features of the UKC framework. The UKC already integrated the entire EOMW content, including Gaelic. As both the EOMW synsets and the translations are aligned with PWN synsets, technically the merge operation was straightforward.

The importing of translations into the UKC was executed in a fully monotonic way: if a newly translated word lexicalised an existing synset then it was added to it as a synonym, while if no such synset existed in the UKC then it was created on the fly. During this process we also stored the provenance of each synset and word (EOMW or the translator). Whenever the same lexicalisation was provided by both sources, they were fused but the provenance indicated both sources, a fact that we exploited in our validations and statistics.

This automated merge operation did not take into account the same word being provided by both resources but with differing orthographies. This is a relevant issue for minority languages in general, as words in such languages often do not have canonical orthographies prescribed by an authoritative source, or have more than one of them. For this reason, based on local dialects, the same word is often spelled in multiple ways. We have chosen not to eliminate such duplicates, as the presence of multiple acceptable orthographies contributes to the richness of the resource.

As the UKC formally represents phenomena of lexical diversity (Giunchiglia et al., 2017), we also imported the lexical gaps marked up by the translator along with the regular synsets. These were formalised in the UKC as special synsets with a gloss but without a lexicalisation.

Finally, the export facility of the UKC was used to output the merged wordnet as a single spreadsheet that served as a basis for validation.

Translations	%	Neologisms	%
Correct	94.3	Correct and new	75.5
Incorrect	4.7	Correct but not new	17.1
Unclear	1.0	Not accepted	7.4
Gaps	**%**	**Glosses**	**%**
Confirmed	90.5	Correct	n/a
Word exists	9.5	Incorrect	n/a

Table 1: Validation results on translations (already existing Gaelic words), neologisms, gaps, and glosses (still under validation).

	USGW	Transl.	EOMW
Words	**10,187**	6,459	4,371
Senses	**15,143**	8,969	6,657
–Noun	**12,181**	7,487	4,780
–Verb	**1,872**	1,097	777
–Adjective	**948**	9	939
–Adverb	**176**	14	162
Synsets	**13,617**	8,911	5,132
Glosses*	**8,030**	8,030	0
Examples*	**265**	265	0
Lexical gaps	**664**	664	0

Table 2: Statistics on the (merged and validated) Unified Scottish Gaelic Wordnet, as well as on its two sources (* = still under validation).

4.4. Validation

Our validation method explicitly and formally addressed individual word translations and their quality, as well as neologisms and lexical gaps. It also considered in a non-exhaustive manner the translations of glosses.

There were two reasons for validation happening after merging. Firstly, it allowed the two resources to validate each other through 'inter-wordnet agreement': we excluded word senses that overlapped between the two sources (the same word expressing the same meaning), automatically considering them as correct. There were 362 such cases of overlap (3.5% of all translations). Secondly, it allowed a global qualitative evaluation of the entire merged wordnet. Beyond this global overview, we purposefully excluded the Wiktionary-derived words from word-by-word validation because that resource has already been evaluated in (Bond and Foster, 2013a) and its contents are not part of our contribution.

The validation was carried out by the fifth author, a native speaker and language expert with experience in professional translation tasks for the Scottish Government. She was contracted to carry out the following tasks:

- *translated words:* validate the correctness of all translated words by marking them up as *correct*, *incorrect*, or *unclear* for borderline cases, and by providing alternative translations for incorrect ones;
- *neologisms:* validate all proposed neologisms by marking them up as *correct*, *correct but not new* (in case the supposedly new word or expression already existed), or *not accepted* (in case another Gaelic word already existed to express the meaning or the validator did not consider it as a desirable suggestion for any other reason);
- *gaps:* validate the meanings marked as lexical gaps by the translator, either as confirmed gaps or as non-gaps due to an existing lexicalisation, which the validator needs to indicate;
- *glosses:* provide a fast and possibly non-exhaustive validation of glosses, addressing only evident mistakes and omissions;
- *global overview:* provide a global, qualitative overview of the merged wordnet resource as a whole.

In total, validation addressed all 8,969 translated word senses, including 2,393 neologisms (that is, 26.7% of all translations provided!). The results can be seen in Table 1. On existing words, the correctness was found to be 94.3% while on neologisms it was 92.6% (although not all of the neologisms deemed correct were found to be actually new, as shown in table 1), and on lexical gaps it was 90.5%. The overall before-validation accuracy was thus 93.1% (again, excluding the EOMW entries that were not evaluated). We consider these results to be a strong evidence of the high quality of the translator's work. Considering that for each incorrect translation the validator provided a correct alternative, the correctness of the final result is very close to 100% (assuming the validator's suggestions to be correct).

Globally and qualitatively, the validator found the new translations to be of higher register (i.e. more formal) then the Wiktionary entries. Some of them, while technically correct, were deemed less recognisable to speakers than the Wiktionary equivalents. The validator also pointed out that additional synonyms could still be provided for many of the translated synsets. For translations considered as mistakes, we gathered the following statistics from the validator's comments: 68% of the mistakes were spelling mistakes, 13% typos, 3% words that were deemed too rare with a much more common alternative available, and the rest (16%) disagreements on the meanings of specific words.

5. Results, Statistics, and Discussion

Table 2 contains statistics on the final, merged and validated wordnet resource.

In order to give an impression of the size of the resource with respect to other wordnets, we have computed the rank of our wordnet with respect to all other wordnets found in the EOMW (which incorporates most wordnets in the world). By the number of words, our wordnet ranked 30th while by (English-aligned) synsets it was the 25th largest resource, among over 1,000 wordnets.

Table 3 provides additional insights into the properties of the wordnet, and also compares it to Princeton WordNet, the most complete such resource. Thus, our USGW has its fair share of verbs but is relatively poor in adjectives with respect to the PWN (that has 15% of adjectives), which should be addressed in future work. The *average polysemy* (how many meanings a word has on average) of USGW is 1.49, slightly higher than that of PWN (1.33). This is

	USGW	Transl.	EOMW	PWN
Avg polysemy	**1.49**	1.37	1.52	1.33
Avg synonyms	**1.17**	1.08	1.30	1.76
Noun senses	**80%**	83%	72%	70%
Verb senses	**12%**	17%	12%	12%
Adj. senses	**6%**	0%	14%	15%
Adv. senses	**1%**	0%	2%	3%

Table 3: Comparison of statistics computed on the Unified Scottish Gaelic Wordnet, its source components, and Princeton WordNet 2.1.

probably due to the incomplete coverage of predominantly monosemous specialised terminology contained in PWN. The average number of synonyms (how many lexicalisations per meaning) is 1.17, much lower than in PWN (1.76). This is partly due to our translator having added a low number of synonyms to each meaning (the value is a mere 1.08 for the translated content).

After validation, the final number of neologisms is 1,807, which still amounts to 20% of the translated content and to 12% of the entire USGW. We hope that this significant amount of new content might be beneficial for the continued use of the Gaelic language.

The merged wordnet resource is described and is downloadable from the web.[15] For the first published version, we used the triple-based file format of the Open Multilingual Wordnet, where each triple describes a synset:

$$(princetonSynsetId, property, value).$$

However, we extended the triples into quadruples, the fourth element indicating the provenance of the information within the merged wordnet (either EOMW or our own effort):

$$(princetonSynsetId, property, value, provenance).$$

A second file published contains 664 validated Gaelic lexical gaps. The format used is exactly the same as for lexicalisations, the only difference being that for gaps the lemma indicated is always the string 'GAP'.

A third file contains another 73 lexical gaps, this time in English: these Gaelic words without English equivalents were provided by our validator. Examples of Gaelic-specific words include *'onfhadh'* (meaning *the raging sound of the sea*) and *'turadh'* (meaning *when the rain stops*). As these meanings have no equivalent in PWN, they are not associated with any existing synset ID. Nevertheless, they are incorporated into the UKC database as new, Gaelic-specific concepts.

In the future we intend to publish the resource in other standard wordnet formats that allow for greater expressivity (e.g. in terms of metadata).

6. Conclusions and Future Work

We consider the Unified Scottish Gaelic Wordnet to be a significant addition to existing Gaelic language resources: it mostly contains original content of very high quality, it is made available for free download and use, it is sense-aligned with the wordnets of over 1,000 other languages of the world, and it is computer-processable, allowing its exploitation for natural language understanding or any other application on Gaelic.

As future work, we foresee the need to extend the coverage of adjectives and adverbs within the resource, which are currently rather weakly covered. We also wist to extend the list of (currently 73) Gaelic-specific (or Celtic, etc.) concepts and the corresponding English lexical gaps. Such gaps cannot be discovered through an English-to-Gaelic translation-based approach and have to be collected using different approaches. We foresee the identification of such concepts, their formal representation, and their integration into the UKC resource to be part of our future research on language diversity (Giunchiglia et al., 2018).

[15] http://ukc.disi.unitn.it/index.php/gaelic/

Acknowledgements

This research was funded by the University of Edinburgh through the DReaM Group EPSRC Platform Grant EP/N014758/1, as well as by the University of Trento through the InteropEHRate project. InteropEHRate is funded by the European Union's Horizon2020 Research and Innovation programme under grant agreement number 826106.

7. Bibliographical References

Agirre, E. and Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Armstrong, R. (1825). *A Gaelic Dictionary: In Two Parts I. Guelic and English. - II. English and Gaelic*. Number v. 1. Duncan.

Batsuren, K., Bella, G., and Giunchiglia, F. (2019). Cognet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145.

Bella, G., Zamboni, A., and Giunchiglia, F. (2016). Domain-based sense disambiguation in multilingual structured data. In *The DIVERSITY Workshop at ECAI 2016, The Hague, The Netherlands*, page 53.

Bond, F. and Foster, R. (2013a). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Bond, F. and Foster, R. (2013b). Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. Association for Computational Linguistics.

Census. (2011). Scotland's census 2011: Gaelic report (part 1).

Dwelley, E. (1990). *The Illustrated Gaelic-English Dictionary*. French & European Publications, Incorporated.

Ganbold, A., Chagnaa, A., and Bella, G. (2018). Using crowd agreement for wordnet localization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Giunchiglia, F., Batsuren, K., and Bella, G. (2017). Understanding and exploiting language diversity. In *IJCAI*, pages 4009–4017.

Giunchiglia, F., Batsuren, K., and Freihat, A. A. (2018). One world–seven thousand languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 18–24.

Kornai, A. (2013). Digital language death. *PloS one*, 8(10).

Maclachlan, E. (1828). *Dictionarium Scoto-Celticum: A Dictionary of the Gaelic Language : Comprising an Ample Vocabulary of Gaelic Words, as Preserved in Vernacular Speech, Manuscripts, Or Printed Works, with Their Signification and Various Meanings in English and Latin, Illustrated by Suitable Examples and Phrases, and with Ethymological Remarks, and Vocabularies of Latin and English Words, with Their Translation Into Gaelic. 1*. Blackwood.

McLeod, W., (2013). *'Chan eil e even ann an Dwelly's!': The Continuing Legacy of Edward Dwelly's Gaelic Dictionary*, pages 163–170. Shaker Publishing.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

Ó Donnaíle, C. (2014). Tools facilitating better use of online dictionaries: Technical aspects of multidict, wordlink and clilstore. In *Proceedings of the First Celtic Language Technology Workshop*, pages 18–27. Association for Computational Linguistics, 8.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142.

Rosch, E. (1999). Principles of categorization. *Concepts: core readings*, 189.

Vossen, P. (1998). Eurowordnet: A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.