lemonGAWN: WordNet Gaeilge as Linked Data

Jim O'Regan, Kevin Scannell, Elaine Uí Dhonnchadha

Trinity College Dublin, Ireland
Saint Louis University, Missouri, USA
Trinity College Dublin, Ireland
jaoregan@tcd.ie, kscanne@gmail.com, uidhonne@tcd.ie

Abstract

We introduce *lemonGAWN*, a conversion of WordNet Gaeilge, a wordnet for the Irish language, with synset relations projected from EuroWordNet. *lemonGAWN* is linked to the English WordNet, as well as the wordnets for four Iberian languages covered by Multilingual Central Repository (MCR), and additionally contains links to both the Irish and English editions of DBpedia.

1. Introduction

WordNet (Miller et al., 1990; Fellbaum, 1998) is a lexical database for English. It relates words to *lexical senses*, which represent different senses of those words, and groups those senses into *synsets*, and provides sets of relations between these synsets (additionally, a number of lexical relations are provided between senses). The synsets and their relations form a semantic graph of English.

Initially developed at Princeton to model psycholinguistic theories of human lexical memory, it has found uses in a number of areas, including various areas of natural language processing; its usefulness in several areas have lead to the creation of wordnets for several other languages, such as the wordnets in the EuroWordNet project (Vossen, 1998), typically linked to Princeton WordNet (PWN). Through these links, the relations can be projected, applying the (largely language independent[1]) semantic graph to the other language; the other uses of wordnet aside, this semantic graph, when connected to the lexical units of a language, is a valuable linguistic resource.

WordNet Gaeilge (described in section 2.) is a wordnet for Irish *(Gaeilge)*, linked to PWN. To make the data more accessible, we are making it available as linked data (section 3.); in addition to providing a pre-generated version of PWN's semantic graph, applied to Irish, we provide links to a number of other wordnets.

2. WordNet Gaeilge and LSG

WordNet Gaeilge is based on *Líonra Séimeantach na Gaeilge* (LSG), an Irish wordnet originally created in 2006 by Scannell[2]. The synsets in the LSG map to PWN synsets in a two-step process. The first step uses English "glosses" in the lexical database[3]. Where the English glosses are unambiguous, they are mapped directly: *stáplóir* is glossed as "stapler" and this lies in a unique PWN synset.

The second step disambiguates the remaining glosses using a sentence-aligned corpus of English texts and their Irish translations, for example, the word *bruach* which has "bank" as one of its glosses. Irish sentences containing *bruach* (or inflected forms, such as *bhruach*, *mbruach*, etc.) are extracted along with the corresponding English sentences. Some of the English sentences will contain the word "bank", and the additional context provided by these sentences is used to decide which is the correct sense of "bank" using standard techniques in word-sense disambiguation. To ensure enough data are available, the bilingual corpus is quite inclusive: the Irish words and their glosses are included, even though they do not form complete sentences, as the glosses alone are often sufficient to determine the correct sense. This fact is well-known to lexicographers, including Ó Dónaill (1977), who gloss words like *feileastram* with two ambiguous English words ("flag, iris") but with no fear of confusion.

Wordnet Gaeilge synsets	77814
Missing from PWN	28356
Missing Irish label	5936
Nouns	49889
Verbs	11548
Adjectives	15250
Adverbs	1127

Table 1: WordNet Gaeilge synsets

WordNet Gaeilge does not link directly to PWN synets, instead using the "sense keys" which identify lexical units, because for many words, the sense distinctions made by the Princeton lexicographers are too fine even for intelligent non-lexicographers to make reliably, and certainly too fine for statistical methods. In addition, there are many distinctions made in Irish that are not made in English (e.g. *dearg* ("red") vs. *rua* ("red", in reference to hair) in Irish would map to a single Princeton synset) and these are precisely the distinctions one does not want to give up in a monolingual Irish language resource. For these reasons, a separate layer – an "intermediate wordnet" – was added between Irish and English, with mappings in both directions. It's still really an English wordnet, but one that is tailored to the needs of Irish. de Bhaldraithe (1959) was very useful in constructing this; the senses given under each English word give a rough first approximation of the sense inventory of the intermediate wordnet.

[1] EuroWordNet, and other similar wordnet projects, use a set of relations that are modified specifically for language independence.

[2] See http://borel.slu.edu/lsg/

[3] These are usually one- or two-word definitions like those found in Ó Dónaill (1977).

LSG is developed as an Open Source project[4], available under the terms of the GNU Free Documentation License, as are the resources described in the present paper. Table 1 gives the current status[5] of synsets in WordNet Gaeilge.

3. Linked Data

Linked Data (Bizer et al., 2009) builds on the technologies of the Web, such as URIs and the HTTP protocol, as a means of creating typed links between data from different sources, using RDF (Resource Description Framework)[6]. The W3C outlines four rules for making data available:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

(Berners-Lee, 2011)

Linked Data has been embraced in recent years by creators of linguistic data as a means of overcoming various problems relating to the inter-operability of disparate data sources Chiarcos et al. (2012). *lemon* (McCrae et al., 2012) describes a model for "ontology lexica" that, among other lexical resources, is being used as the basis for several RDF conversions of wordnets, such as the Chinese WordNet (Lee and Hsieh, 2015).

3.1. Wordnets as Linked Data

PWN, in various versions, has been converted to RDF several times. van Assem et al. (2006) describes a conversion of PWN 2.0, provided by W3C, which has served a central role in the Semantic Web for several years. In our conversion of WordNet Gaeilge, we considered three conversions:

- McCrae et al. (2014) describe a version of PWN 3.1 using the *lemon* model. As the data is hosted by Princeton, this can be considered the canonical version.
- VU Amsterdam provide a conversion of PWN 3.0[7], following the W3C model van Assem et al. (2006).
- IULA Universitat Pompeu Fabra provide a *lemon*-based conversion[8] of Multilingual Central Repository (MCR) (González et al., 2012), a EuroWordNet-based collection of wordnets, including English (based on PWN 3.0), Spanish, Basque, Catalan, and Galician.

As Princeton's RDF is based on PWN 3.1, and not version 3.0 used as a basis for WordNet Gaeilge, we chose not to use it as the basis for our conversion at the present time: each new edition of PWN adds, merges, splits, and deletes synsets, and mapping between versions is non-trivial[9]. However, as it implements the W3C's Best Practices for Converting WordNets to Linked Data[10], we followed its example in how to model our conversion.

As Princeton's RDF is the canonical edition, we wish to both introduce links to it. Although a set of synset mappings from PWN 3.0 to PWN 3.1 is available, as WordNet Gaeilge is based primarily around word senses, not synsets, this is not a complete solution; on the other hand, updating the sense links affects the WordNet Gaeilge database, while a guiding goal in our RDF conversion was to not alter the underlying data.

The RDF conversion of WordNet Gaeilge is being performed in the context of a project to create an Irish-English machine translation system, based on the Apertium platform (Forcada et al., 2011). The Multilingual Central Repository (MCR), as it is based on EuroWordNet, includes links to EuroWordNet's Top Ontology (Vossen et al., 1998), which include classifications of nouns that are necessary to Irish-English translation, in particular "Human" (in English to Irish translation, to select the correct numeral) and "Occupation" (in Irish to English translation, to disambiguate between a subject location *(x is in his house)* and a subject attribute *(X is a teacher)*, which share the same syntactic structure in Irish, but have a different semantic structure in English, e.g., *tá sé ina theach* ("he is in his house") and *tá sé ina mhúinteoir* ("he is a teacher")[11]). For this reason, we chose MCR as the basis for our projection of synset relations. Table 2 contains the number of relations obtained through projection.

VU Amsterdam's conversion follows the W3C's edition of PWN 2.0, and uses the PWN-specific model of that edition. McCrae et al. (2014) make the case that *lemon*'s open model is more suited for interlinking with other resources, so we chose not to use it for our primary conversion. On the other hand, the closed model is more amenable to testing for constraint violations, so our scripts generate a second conversion following this model. In addition, the VU Amsterdam conversion includes semantic relations (Fellbaum et al., 2009) which are not available in other conversions, that further classify the derivational relations in PWN.

3.2. Linking to other resources

Our primary targets in generating links to other datasets have been to other wordnets, currently, VU and the five languages available as part of MCR. In addition, a number of other lexical resources include their own conversions of PWN, such as Lexvo (de Melo, 2013) and Uby (Gurevych

[4] https://github.com/kscanne/wordnet-gaeilge

[5] Git revision a7078d6173735107e838938b3a36360a4da6f9a7, 2015-09-10.

[6] http://www.w3.org/TR/rdf-concepts/

[7] http://datahub.io/dataset/vu-wordnet

[8] https://github.com/martavillegas/EuroWordNetLemon

[9] The website for Princeton's RDF claims to include synset identifiers from PWN 2.0 and 3.0, but at the time of writing, these were not functional.

[10] https://www.w3.org/community/bpmlod/wiki/Converting_WordNets_to_Linked_Data

[11] A further ambiguity exists with states, but it is less easily resolved using wordnet data.

has_hyperonym	14965
has_hyponym	14965
has_mero_madeof	151
has_mero_member	295
has_mero_part	1617
has_subevent	155
is_caused_by	55
is_derived_from	360
is_subevent_of	155
near_antonym	1818
near_synonym	2946
region	155
region_term	155
related_to	16590
see_also_wn15	1684
state_of	472
sumo_at	1364
sumo_equal	749
sumo_plus	25503
topConcept	79757
usage	160
usage_term	160
verb_group	250

Table 2: Synset relations obtained by projection from EuroWordNet (MCR).

et al., 2012), via its RDF conversion, lemonUby (Eckle-Kohler et al., 2014). Table 3 contains the number of sense links to other data sets, table 4 contains the number of synset links.

DBpedia (Auer et al., 2007), an effort to extract Linked Data from Wikipedia, has emerged as a central hub for Linked Data, due to the broad topic coverage of the underlying data. As well as the data from the English edition of Wikipedia, a number of internationalization chapters have been set up, to extract DBpedia data from various language editions of Wikipedia. A chapter for Irish[12] has been set up, though is not currently hosting the extracted data (that is, although data for Irish is available, the URIs it contains are not currently available via HTTP). We provide links to the Irish edition, in anticipation of their availability; we additionally provide links to the English edition of DBpedia, to be immediately useful.

The links to DBpedia were primarily generated via a set of mappings from Google's (now defunct) FreeBase[13]. Although these links were validated by humans, by presenting the PWN gloss and the Wikipedia page related to the FreeBase topic, a number of errors have crept in. In addition to that, the links date from 2012, and do not reflect changes made in either FreeBase or Wikipedia. As part of closing down FreeBase, Google made their data available to the Wikidata project; we plan to regenerate our links based on those which have been validated by Wikidata contributors as the data becomes available.

Logainm (Grant et al., 2013) is a bilingual Linked Data resource for Irish placenames. At present, we only have 5 links from WordNet Gaeilge to Logainm. This is perhaps due to the relatively low coverage of Irish placenames in PWN. We plan to investigate if further links can be found in the Irish-specific synsets.

VU PWN 3.0	97639
Lexvo	27254

Table 3: Word sense links to other data sets

VU PWN 3.0	37607
MCR (English)	97639 (34116)
MCR (Basque)	97639 (34116)
MCR (Catalan)	97639 (34116)
MCR (Spanish)	97639 (34116)
MCR (Galician)	97639 (34116)
lemonUby	37743
DBpedia (en)	7167
DBpedia (ga)	2197
Logainm	5

Table 4: Synset links to other data sets. The number of unique synsets, where relevant, is given in brackets.

4. Future work

The primary goal of future work around *lemonGAWN* is in making it available. Although it is planned to make the data available as "5-star Linked Open Data" (Berners-Lee, 2011), practical concerns have delayed this; in the meantime, the data is being provided via Github[14] under the same terms as WordNet Gaeilge. In addition, scripts used in preparing the data are also being made available[15].

EuroWordNet contains a number of synsets not present in PWN, as does WordNet Gaeilge. As there is likely to be overlap between these synsets, future work will focus on introducing links between them: where an English label is available in both wordnets, we will use the method outlined in section 2.; for the remainder, we will investigate using the other lexical resources available via Lexvo and lemonUby in a similar manner.

SentiWordNet (Baccianella et al., 2010) extends PWN for use in opinion mining and sentiment analysis, by attaching sentiment scores to each synset. We have projected these scores, in the same manner that other links were projected. Ongoing work aims at validating the resulting scores, for use as a sentiment analysis lexicon for Irish. Work on creating chatbots for Irish (Ní Chiaráin and Ní Chasaide, 2016) is incorporating this sentiment analysis lexicon.

Although much work on building wordnets focuses on synset-level relationships, PWN additionally provides lexical links, which provide more fine-grained information about particular words, such as derivational relationships, or connecting verbs to their particles. As a pilot for adding

[12] http://ga.dbpedia.org/

[13] Downloaded from https://code.google.com/p/mlode/downloads/list.

[14] https://github.com/jimregan/lemonGAWN.

[15] https://github.com/jimregan/gawnrdf.

derivational relationships to our conversion, we have focused on adding lexical antonyms; by selecting words from each synset and the antonymic synset, and checking for common prefixes indicative of negation, we have added an initial set of 275 lexical antonyms. Ongoing work in this area concentrates on extracting other derivational relationships, using pairs of affixes across WordNet Gaeilge and PWN, while future work will aim at extracting non-derivational lexical relationships, using corpus-based methods, based on the observation that words collocate with their antonyms.

References

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives, 2007. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudr-Mauroux (eds.), *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 722–735.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani, 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC 2010*, volume 10.

Berners-Lee, Tim, 2011. Linked data-design issues (2006). http://www.w3.org/DesignIssues/LinkedData.html. [Accessed September 16th, 2015].

Bizer, Christian, Tom Heath, and Tim Berners-Lee, 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann (eds.), 2012. *Linked Data in Linguistics*. Springer.

de Bhaldraithe, Tomás, 1959. *English-Irish Dictionary: With Terminological Additions and Corrections*. An Gúm.

de Melo, Gerard, 2013. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web*:7.

Eckle-Kohler, Judith, John P. McCrae, and Christian Chiarcos, 2014. lemonUby–a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, submitted. special issue on Multilingual Linked Open Data.*

Fellbaum, Christiane (ed.), 1998. *WordNet: An Electronic Lexical Database*. Wiley Online Library.

Fellbaum, Christiane, Anne Osherson, and Peter E. Clark, 2009. Putting Semantics into WordNets "Morphosemantic" Links. In Zygmunt Vetulani and Hans Uszkoreit (eds.), *Human Language Technology. Challenges of the Information Society*, volume 5603 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 350–358.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers, 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

González, Aitor, Egoitz Laparra, and German Rigau, 2012. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC'12)*.

Grant, Rebecca, Nuno Lopes, and Catherine Ryan, 2013. Report on the Linked Logainm project. Technical report, Dublin: Royal Irish Academy and National Library of Ireland; Galway: NUI Galway.

Gurevych, Iryna, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth, 2012. Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lee, Chih-Yao and Shu-Kai Hsieh, 2015. Linguistic Linked Data in Chinese: The Case of Chinese Wordnet. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*. Beijing, China: ACL.

McCrae, John, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al., 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

McCrae, John P., Christiane Fellbaum, and Philipp Cimiano, 2014. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*. Reykjavik, Iceland: ELRA.

Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller, 1990. Introduction to WordNet: An On-Line Lexical Database. *International journal of lexicography*, 3(4):235–244.

Ní Chiaráin, Neasa and Ailbhe Ní Chasaide, 2016. Chatbot technology with synthetic voices in the acquisition of an endangered language: motivation, development and evaluation of a platform for Irish. In *Proceedings of LREC 2016 (to appear)*.

Ó Dónaill, Niall, 1977. *Foclóir Gaeilge-Béarla*. Oifig an tSoláthair.

van Assem, Mark, Aldo Gangemi, and Guus Schreiber, 2006. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06), Genoa, Italy*.

Vossen, Piek, 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. *Computers and the Humanities*, 32(2-3).

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters, 1998. The EuroWordNet Base Concepts and Top Ontology. Technical report, Paris, France.

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.