*L*inguistic *I*ssues in *L*anguage *T*echnology – *LiLT*
Volume 10, Issue 4 Sep 2017

History and Development of the Irish Language Semantic Network

Special Issue on Linking, Integrating and Extending Wordnets

Kevin P. Scannell
Jim O'Regan

Published by CSLI Publications

History and Development of the Irish Language Semantic Network

Special Issue on Linking, Integrating and Extending Wordnets

Kevin P. Scannell, *Saint Louis University, St. Louis, Missouri,*
Jim O'Regan, *Trinity College, Dublin, Ireland*

1.1 Introduction

Líonra Séimeantach na Gaeilge (the Irish language semantic network), or LSG for short,[1] is a wordnet for Irish developed originally by Scannell in 2006–2007.[2] With almost 40,000 lexemes organized into 34,000+ synsets, it provides broad coverage of the modern Irish language, and is available under an open source license. The great majority of synsets in LSG are mapped to synsets in the Princeton WordNet (PWN), although we have recently embarked on a project to add synsets corresponding to concepts not lexicalized in English (and therefore not available in PWN); more on this below in §1.2.2.

This paper will give an overview of the project, from the its original conception in the early 2000's, through its construction and initial release in 2007, to the most recent releases which make the database available in a number of linked data formats (O'Regan et al., 2016).

[1] There is a definite article already embedded in the Irish name, and so it is best to refer to it without a definite article in English: "LSG was developed...", etc.

[2] See `https://cadhan.com/lsg/`.

LiLT Volume 10, Issue 4, Sep 2017.
History and Development of the Irish Language Semantic Network.

We will begin with an overview of the Irish language and the state of NLP technology for Irish in §1.1.1. We then cover the history of LSG in §1.1.2, including some information on the method of construction. In §1.1.3, we discuss several end-user resources based on LSG. This is followed by some information on the current state of the project in §1.2, together with statistics on the number of words, word senses, and synsets in the latest version. We close with some of our plans for future development of LSG in §1.3.

1.1.1 The Irish Language

Irish is one of the six extant Celtic languages, very closely related to Scottish and Manx Gaelic, and a bit farther from Welsh, Cornish, and Breton. It is the first official language of the Republic of Ireland, and one of 24 official languages of the European Union. Once spoken throughout the island of Ireland, it is now spoken as a community language only in small *Gaeltacht* (Irish-speaking) regions.

In the most recent (2016) census in the Republic of Ireland, 1.76 million people said they were able to speak Irish (39.8% of respondents), although just 73,803 (1.7% of people age 3 or over) reported that they speak the language daily outside of the education system, and only 20,586 of these people reside in the remaining *Gaeltacht* areas.

By any measure, Irish is an endangered language, and this provides important context for the work described here, and the authors' efforts more generally to develop useful language technologies for those attempting to conduct their lives through the medium of Irish.

Irish is relatively well-resourced in terms of NLP technologies (Judge et al., 2012), at least in comparison with most European minority languages and with the indigenous languages of Africa, Australia, and the Americas. Irish speakers benefit from spelling and grammar checkers, speech synthesizers, machine translation engines, online dictionaries and terminology databases, to name just a few end-user applications. In terms of underlying NLP technologies, there is of course the semantic network that we will describe in the next section, high-quality part-of-speech taggers (Uí Dhonnchadha and van Genabith, 2006, Lynn et al., 2015), and parsers trained on a dependency treebank (Lynn, 2016).

1.1.2 History of LSG

In this section we will provide some historical context for the project and describe the method used to construct the initial release of LSG in 2007.

The landscape in terms of language technology for Irish before the year 2000 was rather bleak. There were a few online bilingual glossaries

and collections of unofficial technical terms assembled by language enthusiasts, but not much more than that, not even a comprehensive monolingual wordlist for spellchecking. This is not to say that there were no good resources at all; indeed, there is a long history of high-quality bilingual (English-Irish and Irish-English) lexicography in Ireland, stretching as far back as Ó Beaglaoich and MacCurtin (1732), and continuing into the 20th century with the publication of landmark English-Irish (de Bhaldraithe, 1959) and Irish-English (Ó Dónaill, 1977) dictionaries, along with a number of specialized terminology dictionaries. Unfortunately, at the time, none of this material was available in electronic form.[3]

Starting in around 1997, Scannell began the development of a fully-featured lexical database of Irish, which has evolved over the last twenty years to serve as the backend for a number of important resources: a spell checker, grammar checker, predictive text software, an Irish "standardizer" (software that brings older texts into conformance with a significant spelling reform that occurred in the 1940's and 1950's, see (Scannell, 2014)), machine translation engines from Scottish and Manx Gaelic into Irish, and, eventually, LSG. Having a unified database across all of these projects has been greatly beneficial since additions and corrections to the database are reflected in each deliverable. Some care is needed, of course; a word like ***loistín*** "ear of a pot" is included in the database but is not a word one would want to include in a spellchecker, since it is rarely used, and any appearance in an Irish text is almost certain to be a misspelling of the common word ***lóistín*** "accommodation, lodging". The database is structured to allow certain words to be "hidden" from certain projects.

The current version of the lexical database stores 48,764 Irish lexemes, 11,234 of these being multiword expressions. Each lexeme has one or more English definitions much like those one would find in a typical bilingual dictionary, but disambiguated with a parenthetical when the English definition might be ambiguous:

```
feileastram nm1: iris (plant), wild iris, flag (iris), yellow
iris, yellow flag, fleur-de-lis (iris).
```

These parentheticals play an important role in what follows. The inventory of possible parentheticals for a given English word is fixed (which is to say that when adding an English definition which is ambiguous to an Irish word, one must choose from a fixed set of possibilities with well-defined meanings; this is **not** to say that the inventory is fixed

[3]We're happy to report that as of this writing in 2017, virtually all of it is; see `http://teanglann.ie/`, `http://focloir.ie/`, and `http://tearma.ie/` for starters.

forever – when gaps are discovered one can add new parentheticals to an English word as needed). This allowed us to link Irish lexemes with specific English word senses, but in retrospect disambiguation via these parentheticals was not an ideal approach: it duplicated effort by generations of English-language lexicographers in enumerating word senses; it relied on the developer's intuition more than careful lexicography to do this; and, given its ad hoc nature, it did not allow easy linking with other standard sense inventories such as PWN. Indeed, the eventual construction of LSG amounted, more or less, to mapping these English word/parenthetical pairs to PWN synsets.

The original aim of the project was to develop an open-source electronic thesaurus that people writing in Irish could easily consult when writing on the computer. At that time, there were a couple of printed Irish thesauri in existence (Ó Doibhlin, 1998, Mac Cionnaith, 2003), but nothing in electronic form, and nothing with sufficiently broad coverage of the language. Thesauri are particularly important for minority languages like Irish for which the majority of speakers are L2 learners. The richness of the language as it is traditionally spoken in the *Gaeltacht* is in danger of dying out; a thesaurus is one small way of giving L2 speakers access to some of this linguistic richness, enabling them to grow their vocabulary and improve their writing without relying on English translations as a crutch.

Our first attempt at developing a thesaurus in (Scannell, 2003) used the lexical database described above as a starting point, and computed similarity scores between Irish words by making use of a public domain version of Roget's Thesaurus available from Project Gutenberg (based on an out-of-copyright edition from 1913). The structure of the resulting thesaurus mirrored the structure of Roget's: long lists of quasi-synonyms arranged into about 1,000 broad categories. It was therefore often difficult to find the exact word one wanted, and particularly so for language learners without combining it with an Irish-English dictionary such as (Ó Dónaill, 1977). Given the shortcomings manifest in the thesaurus, we discarded this work completely, and embarked on a entirely new approach in 2005 with the aim of producing a true wordnet, fully linked with PWN.

The original construction of LSG used the **expand** approach; each English definition in the backend lexical database (or word/parenthetical pair) was mapped to a PWN synset, or to "NULL" when no suitable synset existed in PWN. We attempted, when possible, to facilitate these mappings by making use of word sense disambiguation techniques and an English-Irish parallel corpus, but in the end the mappings amounted to a massive manual (and solo) effort over a two year period. Every

synset in the original LSG corresponded to exactly one synset in PWN (the set of Irish lexemes having at least one English definition mapping to the corresponding PWN synset), and the synset relations were those which could be carried over. PWN synset relations which are lexical in nature (derivational, participles, pertainyms, etc.) and therefore specific to English were not carried over.

It is worth emphasizing that no attempt was made to translate PWN synsets to Irish, as has been done for some other languages (Lindén and Niemi, 2014). If a suitable lexeme exists in our lexical database, then the PWN synset is carried over to LSG, but if no Irish word exists, then the synset is omitted. This means that when viewed as a graph (say of holonym/hypernym relations), there are critical gaps on the Irish side, which are a result of this construction method. For example, there are a number of high-degree nodes on the English side with no corresponding synset on the Irish side: `pwn-3.0:01507175-n` (***bird genus***), `pwn-3.0:01342529-n` (***animal order***), `pwn-3.0:12998815-n` (***agaric***), `pwn-3.0:13756125-n` (***containerful***), `pwn-3.0:13109733-n` (***flowering tree***), and so on. Perfectly reasonable translations of these synsets exist in Irish, but adding them to the database is left for future work. Despite these gaps, our coverage of the 5,000 core PWN synsets is still better than 80%, as reported in Table 1.

Version 1.0 of LSG, released in October of 2007, contained 32,742 synsets, 36,262 lexemes, and 77,596 individual word senses. The current version (1.1) is about 10% larger, as detailed again in Table 1.

LSG has always been available under an open source license, originally the GNU Free Documentation License, and now CC-BY-SA. When it was released, this was rather unusual; just of the 5 out of 53 wordnets listed on the Global WordNet site were open source and freely downloadable in 2007.[4] We are pleased that the landscape has changed greatly over the last ten years, enabling open linked data projects like OMW and ILI.

Despite the fact that LSG has been around for more than ten years, the present paper is the first describing the network in any detail. The only other sources of useful information are the project web site, launched in 2007,[5] and (O'Regan et al., 2016) which covers an RDF version of LSG.

[4]This is a snapshot of the "Wordnets in the world" page on the date LSG was released from the Internet Archive: `http://web.archive.org/web/20071005064800/http://www.globalwordnet.org:80/gwa/wordnet_table.htm`

[5]See `https://cadhan.com/lsg/`.

1.1.3 Applications of LSG

The applications to date of LSG reflect our emphasis on end-user resources; a primary goal of the project has been to make LSG available to Irish speakers in formats that make navigating, searching, and browsing the network as simple as possible.

The first deliverable was a version of the network laid out in book form, freely downloadable as a PDF (Scannell, 2007). The Irish headwords are laid out alphabetically, with numbered senses corresponding to synsets in which the given headword appears, along with cross-references to other synsets when there are holonym/hypernym, meronym/holonym, or other relevant relations. To ease navigation, every word in the PDF is a clickable hyperlink, allowing the user to browse and find the word they want very easily.

We also wanted to make the thesaurus available through an interface tightly integrated with a word processor. For this, we chose OpenOffice.org (and later, LibreOffice) which allows support for user-contributed thesauri via an idiosyncratic file format. While typing a text, the user can click on any word and see a pop-up window showing all related words (which can itself be navigated to follow synset relations, etc.). The main obstacle here was to add support for inflected word forms and link these to the lexemes in LSG, but it was easy enough to achieve this using a suitable Irish stemmer.

As an experiment in visualization of the semantic network, we used an open source 3D graph browser named Morcego[6] to allow one to navigate LSG via a web browser in a visually appealing way. There were some stability and security issues with the interface, so it is no longer online, although some screenshots are still available.[7] Nodes in the graph corresponded to either synsets (colored green) or lexemes (colored red). The lexemes comprising a synset were linked to the green synset node and to each other, and green synsets nodes were linked by an edge for each synset relation. The user was able to use their mouse to rotate the graph in three dimensions, or click on a node to center it in order to explore different parts of the graph.

A final very interesting application of LSG arose out of work by Éilis Uí Mhuirneáin as part of a Master's thesis at the National University of Ireland, Galway (Uí Mhuirneáin, 2010). The focus of the thesis was on the lack of an unabridged monolingual dictionary for Irish, and the difficulties this presents for language learners, teachers, translators, etc.

[6] Morcego is no longer under active development, but the source code is still available from `https://github.com/hacklabr/morcego`

[7] See `https://cadhan.com/lsg/details-en.html`.

TABLE 1 Statistics for Líonra Séimeantach na Gaeilge (lsg-1.1)

Synsets	Words	Senses	Core %	CILI %	Def %	Ex %
34,536	39,625	88,235	82.3	100.0	0.9	0.0

Core % is the percentage of core synsets covered. CILI % is the percentage of synsets linked to CILI. Def and Ex % are the percentages of synsets with definitions and examples respectively

Part of the work involved a small pilot project to assess the feasibility of using LSG to create a monolingual Irish dictionary by providing Irish translations of the PWN glosses linked from LSG. The author (also a professional Irish translator) translated glosses for 300 noun synsets, about 1.3% of the total. These translations have been included in the LMF version of LSG in the `<Definition>` tag for these 300 noun synsets.

1.2 Current state of the LSG

LSG has been under continuous development for ten years since the 1.0 release in 2007. Most of this development has involved additions and corrections to the backend database, often in support of other end-user applications (spelling and grammar correction, MT engines), and as such, incidental to LSG per se. As noted above, these additions have translated into about a 10% increase in the size of LSG in the ten years between version 1.0 and 1.1.

The structure of the network itself and the method of construction have changed very little from the initial version. The only significant changes have been (1) the recent addition of synsets beyond those found in PWN as described below in §1.2.2, and (2) efforts to link LSG with other repositories of open linked data such as DBpedia, Wikidata, the Multilingual Central Repository (MCR), and the Irish placenames database Logainm.ie (O'Regan et al., 2016).

1.2.1 Statistics from the OMW

Statistics for the total number of synsets, words, and word senses in version 1.1 of LSG can be found in Table 1, with a breakdown by part-of-speech in Table 2. LSG contains only a small number of semantic relations above and beyond those already found in PWN; statistics for these are given in Table 3.

1.2.2 Special Characteristics of LSG

The **expand** approach used in the original construction of LSG has obvious limitations, first and foremost that there are many concepts missing from PWN. A good number of these are lexicalized in English

TABLE 2 POS Statistics for Líonra Séimeantach na Gaeilge (lsg-1.1)

POS	Synsets	%	Words	%	Senses	%
Noun	22,339	64.7	26,289	66.3	53,568	60.7
Verb	5,051	14.6	4,943	12.5	13,905	15.8
Adjective	6,741	19.5	7,836	19.8	19,794	22.4
Adverb	405	1.2	557	1.4	968	1.1

TABLE 3 Semantic Relations for Líonra Séimeantach na Gaeilge (lsg-1.1)

Semantic Relation	Count	%
domain region	33	11.9
domain topic	5	1.8
has domain topic	1	0.4
holo member	9	3.2
holo part	38	13.7
hypernym	117	42.1
hyponym	1	0.4
instance hypernym	36	12.9
instance hyponym	32	11.5
mero member	5	1.8
mero part	1	0.4
Constituative	239	86.0
Total	278	100.0

and simply not in PWN (examples are given below in §1.3); others correspond to concepts that are lexicalized in Irish but not in English.

Adding all of the these concepts to LSG would be a massive undertaking; we estimate that at least 5,000 new synsets (and suitable synset relations for each) would be required to cover every Irish lexeme in the current version of the backend database.

As an initial step, we have added about 200 new synsets to LSG, and have proposed these for addition to the CILI. We decided to focus initially on concepts unlikely to be proposed by the maintainers of other wordnets, which fall into two broad categories. First, concepts specific to Ireland or Irish (especially Irish-speaking) culture; for example:

- `lsg-1.1:50000400-n` (***camógaíocht*** "camogie; a field sport very similar to hurling played by women primarily in Ireland")
- `lsg-1.1:50013400-n` (***Galltacht*** "the English-speaking regions in Ireland, as opposed to the Gaeltacht")
- `lsg-1.1:50014900-n` (***lios***, ***ráth*** "a kind of ancient circular fortification found throughout Northern Europe, but especially in Ireland; sometimes called a 'fairy mound' in English")
- `lsg-1.1:50017500-n` (***brídeog*** "an image of St. Bridget used for domestic ceremonies on the eve of that Saint's festival")
- `lsg-1.1:50018900-n` (***fáinneoir*** "one who wears a circular pin showing his or her willingness to speak Irish")

Second, there are many concepts lexicalized in Irish but probably not in any other language, and which are not specific to Ireland or Irish culture. Such words are sometimes called "Dinneenisms", in honor of Fr. Patrick Dinneen, the compiler of a great Irish-English dictionary in the early 20th century that contained many such words (Dinneen, 1927). For example:

- `lsg-1.1:50008000-n` (***bruithneog***, ***gátaire***, ***luathóg***, ***praistéal***, ***prochán***, ***teallachán*** "a batch of potatoes roasted in the ashes of a fire")
- `lsg-1.1:50011400-n` (***méidhe***, "the neck of a headless body")
- `lsg-1.1:50013600-n` (***cothromacan síne***, "the tendency of good and bad spells of weather to offset each other over a period")
- `lsg-1.1:50013900-n` (***codam***, "a swelling of the gums of horses fed on furze")
- `lsg-1.1:50017900-n` (***seicimín***, "the belly skin that falls down between the legs of a well-fed goose")

1.3 Discussion and Future Plans

Although we have begun the process of adding new senses beyond the ones in PWN, there is still much work to do. Of the 48,764 Irish lexemes in the backend database, 6,987 meet our general criteria for inclusion (in terms of part of speech, etc.) yet have no mapping to PWN or any of the augmented senses specific to LSG. Many of the senses represented by these lexemes are, like the ones described above in §1.2.2, unlikely to be lexicalized in other languages. Many others, however, certainly are (even in English), and we expect that dealing these will amount to mapping the missing senses to appropriate CILI entries as they appear. Indeed, many of our gaps are handled already by proposed additions; for example, Irish ***taisire*** "humidifier" maps to `i117741` (from the synset `enwn-3.0:ens-307167`), the technical term ***oistéimiailíteas*** "osteomyelitis" maps to `i117665` (from the synset `odwn-1.3:109823576-n` for ***beenmergontsteking*** in the Open Dutch WordNet), and so on. Others will be straightforward to add: Irish ***maróg ríse*** "rice pudding", ***miotalagrafaíocht*** "metallography", ***Máigheach*** "Mayan" (adjective), ***podchraoladh*** "podcast" (noun), etc.

We are also keen to extend LSG to take advantage of the additional synset relations provided by the ILI framework, in particular relations such as **agent** and **instrument** which will enable linkages between synsets such as `lsg-1.1:10093658-n` (***iascaire*** "one who fishes"), `lsg-1.1:00454121-n` (***iascaireacht*** "the occupation of fishing"), `lsg-1.1:03352628-n` (***eangach*** "fishing net"), and `lsg-1.1:02512053-n` (***iasc*** "fish"), but this remains work in progress.

As mentioned in §1.1.2, lexical relations have not been carried over from PWN. However, the RDF conversion (O'Regan et al., 2016) did include antonym relations, on the basis that the relations were between synsets rather than lexical items, and used a set of derivational affixes to locate lexical antonyms among the lexical items of those synsets, e.g., `lsg-1.1:00004413-a` (***giorraithe*** "abridged") and `lsg-1.1:00004980-a` (***neamhghiorraithe*** "unabridged").

The Irish Terminology Committee (*An Coiste Téarmaíochta*) is a body of experts charged with creating terms for the language; when approved, these are disseminated via an excellent and widely-used website.[8] Using the Terminology Committee's database, we have succeeded in mapping into Irish about 15,000 English terms which only appear in a single PWN synset (which, for this reason, we take to be unambiguous) and which are not already covered by LSG. Many of these are

[8] `http://tearma.ie/`

proper names (***Alfred Dreyfus***, ***Caligula***, ***John Calvin***), and many others are highly technical terms (***amblygonite***, ***clopidogrel bisulfate***, ***endolymph***). Pending approval from the committee, we would like to import these terms wholesale into LSG, which would represent a more than 40% increase in the total number of synsets.

Finally, we note that the majority of the lexicographical work involved in the development of LSG was performed by a single annotator who is an L2 speaker of Irish. While we have received feedback and some corrections from end-users of the wordnet, in the future we would like to involve native speakers in a more formal way to ensure the quality of the synsets and relations.

Acknowledgments

The first author would like to thank everyone who contributed to the original development of LSG in 2005–2007, all of whom are listed on the project website.[9] The second author was funded by the project "Rule-based Machine Translation for Irish-English" from *An Roinn Ealaíon, Oidhreachta, Gnóthaí Réigiúnacha, Tuaithe agus Gaeltachta* under the supervision of Elaine Uí Dhonnchadha at Trinity College, Dublin.

References

de Bhaldraithe, Tomás, ed. 1959. *English-Irish Dictionary*. Baile Átha Cliath: An Gúm.

Dinneen, Patrick S., ed. 1927. *Foclóir Gaedhilge agus Béarla*. Baile Átha Cliath: Irish Texts Society.

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. 2012. *The Irish language in the digital age*. Berlin: Springer.

Lindén, Krister and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation. *Language Resources and Evaluation* 48(2):191–201.

Lynn, Teresa. 2016. *Irish dependency treebanking and parsing*. Ph.D. thesis, Macquarie University.

Lynn, Teresa, Kevin Scannell, and Eimear Maguire. 2015. Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. In *Proceedsings of ACL-IJCNLP 2015*.

Mac Cionnaith, Seán. 2003. *Focail i bhFócas*. Baile Átha Cliath: Coiscéim.

Ó Beaglaoich, Conchobhar and Hugh MacCurtin. 1732. *The English Irish Dictionary. An Focloir Bearla Gaoidheilge*. Paris: Seamus Guerin.

Ó Doibhlin, Breandán. 1998. *Gaoth an Fhocail*. Baile Átha Cliath: Coiscéim.

[9] See `https://cadhan.com/lsg/thanks.html`

Ó Dónaill, Niall, ed. 1977. *Foclóir Gaeilge-Béarla*. Baile Átha Cliath: An Gúm.

O'Regan, Jim, Kevin Scannell, and Elaine Uí Dhonnchadha. 2016. lemonGAWN: WordNet Gaeilge as linked data. In *LDL 2016–5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pages 36–40.

Scannell, Kevin. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the 1st Celtic Language Technology Workshop at COLING 2014*, pages 33–40.

Scannell, Kevin P. 2003. Automatic thesaurus generation for minority languages: an Irish example. In *Actes de la 10e conférence TALN à Batz-sur-Mer*, vol. 2, pages 203–212. ATALA.

Scannell, Kevin P. 2007. Líonra Séimeantach na Gaeilge, 1.001.

Uí Dhonnchadha, Elaine and Josef van Genabith. 2006. A part-of-speech tagger for Irish using finite state morphology and constraint grammar disambiguation. In *Proceedings of LREC 2006, Genoa*.

Uí Mhuirneáin, Éilis. 2010. *Barr do Theanga: Miontráchtas ar Líonra Séimeantach na Gaeilge*. Master's thesis, National University of Ireland, Galway.