Ilfhocail: A Lexicon of Irish MWEs

Abigail Walsh Teresa Lynn Jennifer Foster
ADAPT Centre
School of Computing
Dublin City University
{abigail.walsh,teresa.lynn,jennifer.foster}@adaptcentre.ie

Abstract

This paper describes the categorisation of Irish MWEs, and the construction of the first version of a lexicon of Irish MWEs for NLP purposes (*Ilfhocail*, meaning 'Multiwords'), collected from a number of resources. For the purposes of quality assurance, 530 entries of this lexicon were examined and manually annotated for POS and MWE category.

1 Introduction

Multiword expressions (MWEs), which make up a considerable percentage of our mental lexicon (Jackendoff, 1997), can be a bottleneck in Natural Language Processing (NLP) (Sag et al., 2002). While there are several initiatives dedicated to MWE research – PARSEME (Savary et al., 2017), SIGLEX-MWE Workshops (Savary et al., 2018; Markantonatou et al., 2017; Mitkov et al., 2017) – the focus has tended to be on majority languages (Losnegaard et al., 2016). For many minority languages, a lack of resources has impeded research. Irish is one such minority language. While progress has been made over the past several years in the area of Irish NLP (Uí Dhonnchadha and Van Genabith, 2008; Scannell, 2014; Lynn et al., 2015; Lynn, 2016), there is still a significant lack of technological support for identification and categorisation of MWEs. In fact, as a result, minimal labelling of MWEs is found in both Irish treebanks, Irish Dependency Treebank (Lynn, 2016) and Universal Dependency Treebank (Nivre et al., 2018; Lynn and Foster, 2016).

There have, however, been some theoretical linguistic studies on particular forms of MWEs in Irish. In her analysis of Irish syntax, Stenson (1981) describes idiomatic copular constructions, and verb-object constructions. Bloch-Trojnar (2009) and Bayda (2015) have carried out research on light verb constructions. Ó Domhnalláin and Ó Baoill (1975) have compiled a book of verb-particle constructions and their meanings. A valency dictionary for Irish verbs was created by Wigger (2008) and his team (Foclóir Briathra Gaeilge). Ní Loingsigh (2016) has compiled a database of manually annotated idioms in Irish, taken from the collections of an tAthair Peadair Ó Laoghaire.

Our work aims to compile a comprehensive lexicon of Irish MWEs (*Ilfhocail*) for the purposes of NLP, by leveraging both existing monolingual and bilingual lexical resources and generating new MWE entries through methods of semi-automatic discovery. We compile the data from various sources into a unified structure, and define an MWE categorisation scheme. Our current lexicon contains 201,795 entries and a subset of these will be released, subject to the licensing agreements of the various sources.

We document the design decisions required when combining data from the various lexical sources currently available for Irish (Section 2). We also find that Irish MWEs are not easily categorised according to standard MWE categories (Section 3). We manually examine and categorise a sample of 530 entries, both as a way to evaluate the quality of the extracted MWEs and to assess and inform our categorisation scheme (Section 4).

2 Compiling the lexicon

Although in some respects Irish can be considered a low-resource language, valuable resources in the form of Irish lexicons and Irish-English/English-Irish dictionaries are now available. We extracted MWE entries from the following resources in XML format.

An Bunachar Náisiúnta Téarmaíochta don Ghaeilge (The National Terminology Database

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 162–168
Florence, Italy, August 2, 2019.

for Irish)[1] The Tearma database, consisting of about 185,000 entries, is the largest resource available. 141,031 of these entries were extracted as MWEs, comprising about half of our lexicon. The Tearma database can be downloaded as a *txt* or *tbx* file from https://www.tearma.ie/ioslodail/, and is available for personal use.

Líonra Séimeantach na Gaeilge (Irish Wordnet) This database, created by Kevin Scannell, contains over 32,000 synsets. 8,995 MWE entries were extracted from this resource. It can be downloaded in several formats from https://cadhan.com/lsg/index-en.html under the GNU Free Documentation License.

Peadar Ó Laoghaire Idiom Collection This collection of idioms was extracted from the works of Peadar Ó Laoghaire and annotated with additional information (Ní Loingsigh and Ó Raghallaigh, 2016). All 420 of these entries were added to the *Ilfhocail* lexicon. The searchable corpus is available at https://www.gaois.ie/bnl/en/, and a downloadable version of the corpus was made available to us for research purposes.

Pota Focal Gluais Tí (Pot of Words House Glossary) The House Glossary was created by Michal Boleslav Měchura, and contains over 6,000 terms, 375 of which were extracted as MWEs for the lexicon. It is under the Creative Commons Attribution Non-Commercial Share-Alike licence and can be downloaded from https://github.com/michmech/pota-focal-gluais/.

The New English-Irish Dictionary[1], and the English-Irish Dictionary[1] The electronic searchable version of the English-Irish Dictionary (de Bhaldraithe, 1959) was made available online by Foras na Gaeilge, with their New English-Irish Dictionary released in 2013 with revised entries and additional grammatical information. There were a combined total of 105,358 MWE entries extracted from these dictionaries, though many of these terms were duplicates (see below).

Foclóir Gaeilge-Béarla (Irish-English Dictionary)[1] This is an electronic searchable version of the Irish-English dictionary (Ó Dónaill, 1977). Only 48 of the 59,700 entries were MWEs; however, it was observed that the sense entries contained many idiomatic uses of the entry word. These sense entries (38,775) were added to the our lexicon.

An Foclóir Beag (The Small Dictionary)[1] This is an electronic searchable version of the Foclóir Beag dictionary (Ó Dónaill and Ua Maoileoin, 1991). 771 terms were extracted and added to the lexicon.

2.1 Lexicon Structure

The lexicon is organised under the columns `GA-Head`, `GA`, `POS`, `EN`, `Source` and `ID`. `GA-Head` is the headword of the Irish entry, and corresponds to the word that the entry was filed under. Where this was not available (e.g. in the English-Irish Dictionary, all expressions were under an English headword), the first word of the Irish entry was used. As Irish is a head-initial language, given its VSO word order, and lack of indefinite articles, this was deemed a sufficient default value. The Irish entry was listed under the `GA` column.

While each MWE in the lexicon had an Irish entry, this was not always the case for POS information and English translation, listed under `POS` and `EN` respectively. The POS information extracted from each resource varied from no POS label to broad level POS information (noun, verb, etc.) to more fine-grained syntactic information (transitivity, gender, number, etc.). English translations were present in all resources save the *Líonra Séimeantach na Gaeilge*, the *Peadar Ó Laoghaire Idiom Collection* and the *Foclóir Beag*.

`Source` is a three or four letter string indicating which dictionary it was extracted from. `ID` is a unique string for each entry, created by concatenating the source code with a unique integer.

2.2 Cleaning

Some entries are present in a number of resources and, even within one resource, there are multiple instances of the same Irish MWE, with differing POS or translations. We keep the entries distinct on the POS level (1), but combine MWE entries across different English translations and sources (2). Several of the concatenated English translations for an MWE contain duplicate, redundant information and so any translation that was a substring of another is removed (3).

(1) “Cósta Ríceach”,“**ADJ**”,“Costa Rican”
“Cósta Ríceach”,“**NOUN**”,“Costa Rican”

[1]These resources were provided to us by Foras na Gaeilge for research purposes and are not to be republished

(2) "great and small", "young and old" → "**great and small; young and old**"

(3) "birthday","birthday (Happy Birthday!)" → "birthday (Happy Birthday!)"

Following these steps, the corpus was condensed from 389,424 entries to 201,795 entries.

3 MWE Categorisation

Ideally the lexicon entries would include information about the type of MWE. However, there does not exist an agreed-upon taxonomy of MWEs in Irish to date, although there has been some research investigating certain categories of MWEs, including idioms (Ní Loingsigh, 2016), light verb constructions (Bayda, 2015), verb-particle constructions (Ó Domhnalláin and Ó Baoill, 1975), and other idiosyncratic constructions (Stenson, 1981). Throughout the development of the lexicon, some prospective MWE categories became easily identifiable through the POS tags of their headwords (e.g. Nominal MWEs). Other categories were determined following examples of categorisation efforts in other languages.

In her work on creating a taxonomy of Spanish MWEs, Parra Escartín (2015) describes the various taxonomy schemes of MWEs that have been suggested, such as those of Sag et al. (2002), Baldwin and Kim (2010) and Ramisch (2015). These taxonomies make distinctions between lexicalised phrases and institutionalised phrases. Lexicalised phrases are expressions which are idiosyncratic on a lexical, semantic or syntactic level; institutionalised phrases are considered MWEs based on statistical idiosyncrasy alone.

These taxonomies also distinguish between fixed expressions, semi-fixed expressions and syntactically flexible expressions. We define these terms depending on the variability of the MWE entries as they occur in the manually annotated sample of the lexicon. Fixed expressions do not allow for any variation or inflection, and include fixed idioms such as those listed in section 3.3, as well compound prepositions. Semi-fixed expressions allow some degree of inflection, but the word order is fixed and there are no gaps, e.g. nominal MWEs, some idiomatic constructions with "be". Non-fixed or flexible expressions can be discontinuous, word order may be flexible and elements of the expression may inflect. These expressions include light verb constructions, verb-particle constructions, inherently adpositional verbs and certain idioms.

The initial approach that we take is to broadly categorise Irish MWEs into *non-verbal* and *verbal* MWEs. The categories of verbal MWEs were chosen to align with the PARSEME Annotation Guidelines 1.1 (Ramisch et al., 2018). However, we note that there are a number of MWEs for Irish that do not fall neatly into the PARSEME categories (see section 3.2).

3.1 Non-verbal MWEs

Compound Prepositions Some simple prepositions can combine with a noun to form compound prepositions. These compound prepositions act as fixed lexical items and do not inflect.

(4) *i ndiaidh* 'after'

Nominal MWEs Nominal MWEs (NMWEs) are multiword terms that include named entities, noun-noun compounds, and noun-adjective and noun-prepositional phrase constructions. The majority of the MWE entries in our lexicon appear to be N-N compounds or N-Adj compounds, due in part to the inclusion of the relatively large Tearma database of Irish terminology.

(5) *garrán préachán*
grove of-rooks
'rookery'

3.2 Verbal MWEs

Light Verb Constructions Light Verb Constructions (LVCs) consist of a verb and a noun, the latter of which contributes most of the semantics within the construction. These constructions can be accompanied by a necessary preposition (see Inherently Adpositional Verbs below).

(6) ***Rinne*** *Sorcha* ***iarracht air***.
(make-PA Sarah attempt on-it)
'Sarah tried it.'

Verb-Particle Constructions Verb-Particle Constructions (VPCs) are expressions consisting of a verb and a particle, that is, a preposition or adverb, that changes the meaning of the verb.

(7) *tabhair* 'give'
tabhair amach 'complain'

(8) *buail* 'hit'
buail le 'meet'

The change in the meaning may be significant or subtle.

Inherently Adpositional Verbs Inherently Adpositional Verbs (IAVs) are constructions defined in the PARSEME Annotation Guidelines. These are verb-adposition constructions, where the verb must take a certain adposition.

(9) *maith (rud)* ***do*** *(duine)* 'forgive (something) of (someone)'

This construction does not exactly align with the PARSEME Guidelines, given that the additional adposition occasionally appears to change the meaning of the construction.

(10) *cuir síos* 'put down'
cuir síos ***ar*** 'describe'

It could be argued that the VPC *cuir síos* already allows for this meaning, but never occurs without the adposition in this context.

Idiomatic Constructions with "Be" Irish has two verbs which translate to the English verb "be". The copular verb in Irish (*is*) is used to indicate states, emotions, etc., while the substantive verb (*tá*) is used in periphrastic aspectual constructions (Ó hUallacháin and Ó Murchú, 1981). Both of these verbs are often used in idiomatic constructions (BE-idioms), which function as a unit in Irish.

(11) ***Is maith liom*** *tae.*
(COP good with-me tea)
'I like tea.'

(12) ***Tá áthas orm.***
(be happiness on-me)
'I am happy.'

While we've termed these constructions 'Idiomatic constructions with "be"', they do not align with the PARSEME category of verbal idioms, and are potentially a new category of verbal MWEs.

3.3 Idioms

Idioms as a category of MWE can fall under both verbal and non-verbal MWEs, depending on what the headword is deemed to be. This category allows for expressions that are clearly idiomatic or idiosyncratic, but do not follow a syntactic pattern as described above. They also include various fixed, idiomatic expressions such as proverbs, sayings and non-decomposable expressions.

(13) *Idir dhá thine Bhealtaine*
(between two fire May-GEN)
'Between a rock and a hard place
(lit. between two May fires)'

(14) *Gearraíonn beirt bóthar* 'Easier with two'
(lit. Two shorten the road)

(15) *(a) sheacht míle dícheall* '(his) very best'
(lit. his best seven thousand)

3.4 Institutionalised Phrases

Institutionalised Phrases (IPs) are described in Sag et al. (2002) as expressions that are statistically idiosyncratic. IPs are distinct from collocations in that IPs discount compositional phrases that are predictably frequent for non-linguistic reasons. While these expressions are not idiomatic or non-compositional, their frequency in language creates a strong association between the concept and the expression.

(16) *aire agus forcamás* 'care and attention'

(17) *ceathrar déag* 'fourteen'

Given that the only defining characteristics of institutionalised phrases are their statistical frequency and lack of idiomaticity, distinguishing between IPs and collocations or other lexical chunks that may be included in a dictionary proved challenging when annotating the sample corpus.

4 Manually Annotated Sample

In order to assess the quality of the lexicon, 530 entries were randomly selected from the lexicon and examined. Missing POS and translations were added, erroneous headwords were corrected and the entries were labelled with a MWE category and whether they were fixed expressions (*f*), semi-fixed expressions (*s*) or non-fixed or flexible expressions (*n*). Table 1 demonstrates how the sample MWEs were categorised. The highest proportion of MWEs are Nominal MWEs, mostly originating from the Tearma corpus.

The manual annotation revealed some bugs:

Headwords As mentioned in Section 2, not every resource had headword information, and the default token value assigned to this field was sometimes incorrect. Moreover, there was a lack of consistency in choice of headword across different resources - with some resources choosing headwords of different POS type for different expressions.

Compound Prepositions	2
Nominal MWEs	377
Light Verb Constructions	30
Verb-Particle Constructions	5
Inherently Adpositional Verbs	17
Constructions with 'Be'	2
Idioms	63
Institutional Phrases	31
Non-MWEs	18

Table 1: Categorisation of 530 MWEs

(18) ***an***, *an mhainistir* (the cloister)
headword should be *mhainistir* 'cloister'

(19) ***caobh***, *cara caobh* (gentle friend)
headword should be *cara* 'friend'

POS tags POS information refers to the POS of the headword of the MWE entry. Given how some of the entries did not have a headword, the POS information is lacking for a number of the sources. Moreover, the labels used to denote POS information varies between sources. We aim to unify all these labels for the next release of the corpus.

Non-MWEs There were several instances in the sample that were deemed not to be multiword expressions – see last row of Table 1. These included productive entries, and terms which did not qualify as institutionalised phrases, whether because elements of the expression could be easily replaced by another word (i.e. too productive), or there were too many non-lexicalised components included in the entry.

(20) *súile silteacha* 'streaming eyes'
(*silteacha* is a productive adjective that can be applied to many nouns)

(21) *Dá mbeadh cosúlacht ar bith orthu* 'if they showed any promise'
(Not an idiomatic or statistically idiosyncratic entry)

Productive Entries Many entries, particularly those extracted from the English-Irish Dictionary, included non-lexicalised (i.e. non-core) elements in the expression. As these non-lexicalised elements were often members of a relatively small semantic class of words, it is difficult to decide whether these entries should be considered MWEs. In these contexts, the headword would gain a different meaning.

(22) *gearr* 'cut'
gearr pionós, dualgas, fíneáil 'impose a penalty, duty, fine'

5 Conclusion

We have described the first release of *Ilfhocail*, an Irish MWE lexicon. It was compiled semi-automatically using several lexical resources for Irish, and currently contains 201,795 entries. Issues discovered through manual annotation of 530 entries will be handled in the second version, e.g. unifying POS information, removing non-MWEs and a first attempt at automatic categorisation of MWE type.

A second contribution of this paper is an initial attempt at defining a categorisation scheme for Irish MWEs. This scheme takes categorisation schemes for other languages as a basis and modifies them to accommodate the properties of the Irish language. It is our hope to include Irish in a future version of the PARSEME Shared Task on Automatic Identification of Verbal MWEs. To that end, it is necessary to determine how the categories of verbal MWEs in Irish align to the PARSEME Annotation Guidelines[2], and whether these categories must be modified to fit the annotation scheme or vice versa.

Ilfhocail will serve as a useful source of data for future experiments in Irish NLP. These include automatic identification of MWEs in the Irish Treebanks (Lynn, 2016; Lynn and Foster, 2016), which will facilitate the development of Irish parsing technologies, as well as intelligent MWE handling for improved English-Irish and Irish-English Machine Translation.

Acknowledgements

The first author's work is funded by the Irish Government Department of Culture, Heritage and the Gaeltacht under the GaelTech Project, and is also supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (http://www.adaptcentre.ie) at Dublin City University. The authors would also like to acknowledge Noah Ó Donnaile for his help with the creation of this lexicon, particularly in extracting the entries from the lexical resources.

[2] https://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. *Handbook of natural language processing*, pages 267–285.

Victor Bayda. 2015. Irish constructions with bain. *Yn llawen iawn, yn llawn iaith: Proceedings of the 6th International Colloquium of Societas Celto-Slavica. Vol. 7 of Studia Celto-Slavica. Johnston, D., Parina, E. and Fomin, M. (eds)*, 7:213–228.

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

Maria Bloch-Trojnar. 2009. On the Nominal Status of VNs in Light Verb Constructions in Modern Irish.

R. Jackendoff. 1997. *The Architecture of the Language Faculty*. Linguistic inquiry monographs. MIT Press.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME Survey on MWE Resources. pages 2299–2306.

Teresa Lynn. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University, Macquarie University.

Teresa Lynn and Jennifer Foster. 2016. Universal dependencies for Irish. In *Celtic Language Technology Workshop*, July, pages 79–92, Paris.

Teresa Lynn, Kevin Scannell, and Eimear Maguire. 2015. Minority language twitter: Part-of-speech tagging and analysis of irish tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8, Beijing, China. Association for Computational Linguistics.

Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors. 2017. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, Valencia, Spain.

Ruslan Mitkov, Violeta Seretan, and Gloria Corpas Pastor, editors. 2017. *Proceedings of The 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2017)*. Editions Tradulex, Geneva.

Katie Ní Loingsigh. 2016. *Tiomsú agus Rangú i mBunachar Sonraí ar Chnuasach Nathanna Gaeilge as Saothar Pheadair Uí Laoghaire*. Ph.D. thesis.

Katie Ní Loingsigh and Brian Ó Raghallaigh. 2016. Starting from Scratch – The Creation of Irish-language Idiom Database. pages 726–734.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Kamil Kopacewicz, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko

Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Carla Parra Escartín. 2015. Spanish multiword expressions : Looking for a taxonomy. pages 271–323.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer. https://doi.org/10.1007/978-3-319-09207-2.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. *Computational Linguistics and Intelligent Text Processing*, pages 1–15.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. *Proceedings of The 13th Workshop on Multiword Expressions*, (Mwe):31–47.

Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan, and Miriam R. L. Petruck, editors. 2018. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Kevin Scannell. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Nancy Stenson. 1981. *Studies in Irish syntax*. Ars linguistica. Tübingen: Gunter Narr Verlag.

Elaine Uí Dhonnchadha and Josef Van Genabith. 2008. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. Ph.D. thesis, Dublin City University.

Arndt Wigger. 2008. Advances in the lexicography of Modern Irish verbs. pages 233–250.

Tomás Ó Domhnalláin and Dónall Ó Baoill. 1975. *Réamhfhocail le briathra na Gaeilge*. Tuarascáil taighde. Institiúid Teangeolaíochta Éireann.

Niall Ó Dónaill. 1977. *Foclóir Gailge-Béarla*. An Gúm, An Roinn Oideachas.

Niall Ó Dónaill and Pádraig Ua Maoileoin. 1991. *An Foclóir Beag*. An Gúm, Baile Átha Cliath.

C. Ó hUallacháin and M. Ó Murchú. 1981. *Irish Grammar*. University of Ulster Coleraine.