Annotating MWEs in the Irish UD Treebank

Sarah McGuinness, Jason Phelan, Abigail Walsh and **Teresa Lynn**
ADAPT Centre, School of Computing, Dublin City University, Ireland
sarah.mcguinness26@mail.dcu.ie
jason.phelan09@gmail.com
{abigail.walsh,teresa.lynn}@adaptcentre.ie

Abstract

This paper reports on the analysis and annotation of Multiword Expressions in the Irish Universal Dependency Treebank. We provide a linguistic discussion around decisions on how to appropriately label Irish MWEs using the `compound`, `flat` and `fixed` dependency relation labels within the framework of the Universal Dependencies annotation guidelines. We discuss some nuances of the Irish language that pose challenges for assigning these UD labels and provide this report in support of the Irish UD annotation guidelines. With this we hope to ensure consistency in annotation across the dataset and provide a basis for future MWE annotation for Irish.

1 Introduction

The aim of the Universal Dependencies (UD) project (Nivre et al., 2016) is to facilitate and improve cross-lingual learning and multilingual parsing through the creation of a harmonised set of annotation guidelines for treebanks across multiple languages. As the project and guidelines evolve, new language treebanks are being added at each 6 monthly release. The Irish UD Treebank (IUDT) (Lynn et al., 2017) has been part of the UD project since the v.1 release in 2015, following a conversion from the original Irish Dependency Treebank (IDT)[1] (Lynn, 2016).[2] Until recently, however, there was little opportunity to fully explore the treatment of Multiword Expressions (MWEs) in either the original IDT annotation scheme or the converted UD scheme. This is mainly due to two factors: (i) both treebanks were the product of a PhD dissertation with limited scope, and (ii) prior research (both theoretical and applied) on MWEs in Irish was limited and generally insufficient in terms of Natural Language Processing (NLP) relevance or application (see Section 2). These factors are commonplace as challenges that face low-resource languages such as Irish.

MWEs are reported to make up a large part of natural language, as much as forty percent of our mental lexicon (Jackendoff, 1997; Fellbaum, 1998). As Constant et al. (2017) observe, providing an exact definition of MWEs can be controversial and there are varying interpretations and analyses to be found. These variations may well be due to the differing motivations for the need for definition (theoretical, applied, etc.). Our approach has been informed by the work of Sag et al. (2002), Baldwin and Kim (2010) and Ramisch (2015), whose works lie within the field of NLP. We define MWEs to be a string of two or more tokens, which form a unit at a semantic, syntactic or lexical level.

Research into MWEs, particularly with respect to NLP, has grown substantially since it has become increasingly apparent that they present a bottleneck for automatic processing of human language (Sag et

[1] https://github.com/tlynn747/IrishDependencyTreebank

[2] The data for the v2.6 Irish Universal Dependency Treebank is based on a gold standard POS-tagged sample of the National Corpus of Ireland developed by Uí Dhonnchadha (2009).

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 126–139
Barcelona, Spain (Online), December 13, 2020

al., 2002). In fact, both the ICT COST Action PARSEME (IC1207)[3] and the establishment of the MWE Workshop series[4] were motivated by the need to establish how best to represent and encode MWEs for the benefit of improving NLP across languages.

Syntactic parsing is one particular area that can see improvement in accuracy when additional information is known with respect to the use of Multiword Expressions in a language (Nivre and Nilsson, 2004; Seretan, 2011; Green et al., 2013; Candito and Constant, 2014; Savary et al., 2015). Essentially, a parsing system is expected to perform better when it is aware that a string of words should be treated as one syntactic unit instead of individual tokens (e.g. *They tried to* ***hold up*** *a bank* vs *the container can* ***hold*** *up to 10 gallons*). Given the small size of the Irish UD treebank,[5] it makes sense that both improving the quality of the trees with additional information (such as accurately labelled MWEs) and increasing the size of the dataset should be treated with similar levels of importance.

This paper reports on the labelling of MWEs in the v2.6 release of the Irish UD treebank. A summary of our contribution is as follows: (1) We contribute to the quality of the v2.6 IUDT release by fully reviewing and updating MWE annotation. (2) We propose an approach to analysing and labelling Irish compounds, fixed expressions and flat proper noun strings within the UD framework. (3) In particular, we highlight the issues arising in differentiating between `compound` and nominal modifier (`nmod`) dependents while reporting on a small survey to address this, and hope that this may be helpful to other treebank developers who face similar challenges.

While the UD guidelines aim to capture linguistic universals, there will always be language specific features to consider when meeting the trade-off between cross-lingual consistency and a sufficient representation of that language. Given that this field of research is still relatively under-explored in Irish, some annotation choices may still be controversial to those in the field of Irish linguistics outside the UD framework, but nonetheless this starting point will be useful for any future studies in this area.

2 Related Work

As highlighted by Losnegaard (2016) and Parra Escartín (2018), for a long time, most MWE research was traditionally based on major languages such as English. In terms of linguistic data annotation, analysis of MWEs had only been carried out in a limited number of dependency treebanks (e.g. Czech (Bejček and Straňák, 2010), Hungarian (Vincze et al., 2013) and Turkish (Eryiğit et al., 2015)). In a step to address this, the ICT COST Action PARSEME[6] set out to draw up guidelines for classifying and categorising MWEs across multiple languages. The main goal was to provide a framework for processing MWEs in order to improve performance in the areas of machine translation and parsing (Savary et al., 2015; Savary et al., 2017; Ramisch et al., 2018; Losnegaard et al., 2016).[7]

Complementary work (Rosén et al., 2015; Rosén et al., 2016) proposed general guidelines for annotating MWEs in both constituency and dependency treebanks across 15 languages with the help of a focused survey. Through this, it was observed that it should be possible to search for various types of MWEs based on their characteristics (e.g. compositional vs non-compositional).

In terms of syntactic parsing, Candito and Constant (2014) observe that while MWE information is intuitively supposed to help parsing, it is difficult to prove this in a realistic setting. Difficulties arise when MWEs are automatically identified – an approach which can result in error propagation (Constant et al., 2012). It was also noted that the use of external lexicons did not seem to suffice in MWE processing, and that the use of data-driven external information would potentially help with this. This was supported by Schneider (2014) and the various experiments carried out by Constant et al. (2019) in assessing different approaches to MWE identification and parsing. Constant and Nivre (2016) developed a transition-based system which performed joint syntactic analysis and MWE identification, with promising results based

[3] `https://typo.uni-konstanz.de/parseme/`

[4] `https://www.aclweb.org/anthology/venues/mwe/`

[5] The v2.6 UD release has 2,924 trees and roughly 64,000 tokens.

[6] `https://parsemefr.lis-lab.fr`

[7] The PARSEME guidelines for annotating Verbal MWEs now cover 27 languages `https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/`

on English and French. Overall, the need for MWE-aware treebanks to appropriately train statistical parsers has therefore become more evident.

With respect to UD treebank annotation, the language agnostic guidelines for MWEs deal mainly with compounds, fixed expressions and strings such as proper nouns. The following labels are proposed in the UD v2 guidelines:[8] `compound`; `fixed`; and `flat`. However, as interpretation of MWEs seems to vary across treebank development teams, a working group has been formed in the UD project to seek to harmonise the treatment of MWEs across treebanks.[9] Some treebank groups have also reported on their specific approaches to dealing with MWEs in their data (e.g. Croatian (Sojat and Filko, 2016), English (Schuster and Manning, 2016), Spanish (Martínez Alonso and Zeman, 2016), Turkish (Sulubacak and Eryiğit, 2018) and Farsi (Qasemizadeh, 2014)). More recently, Kahane (2018) provided an alternative proposal for treating idioms and fixed expressions through examining current French and English analyses in UD treebanks.

The following research contributes to our understanding of MWEs in Irish. Stenson (1981) discusses idiomatic constructions including verb-object constructions, verb-particle constructions and idiomatic constructions using the copula. In terms of theoretical studies, Bloch-Trojnar (2009) and Bayda (2015) have both carried out some research on light verb constructions, Ní Loingsigh (2016) on idioms, and Ó Domhnalláin and Ó Baoill (1975) on verbal constructions with prepositions. The Christian Brothers (1999) offer a limited summary of compounds in Irish, but without a specific discussion of MWEs. Uí Dhonnchadha (2009) provides a summary on treatment of phrasal verbs in her work on part-of-speech tagging and a constraint grammar for Irish. The output format of this POS-tagger resulted in some compounds, fixed expressions and proper noun strings being captured in earlier versions of the IUDT (Lynn and Foster, 2016) – the annotation of which were reviewed in this study. In terms of categorisation, a definition of a taxonomy of MWEs in Irish is in its early stages of development (Walsh et al., 2019). This work also discusses phrasal verbs in more detail through the lens of verb particle constructions and inherently adpositional verbs.

Our main sources of reference on Irish grammar in our work are Studies in Irish Syntax (Stenson, 1981), The Christian Brothers' Irish Grammar (Christian-Brothers, 1999) and also online dictionaries such as An Bunachar Náisiúnta Téarmaíochta don Ghaeilge (The National Terminology Database for Irish),[10] The New English-Irish Dictionary,[11] and also the Foclóir Gaeilge-Béarla and An Foclóir Beag which are available at the Teanglann website.[12] A valency dictionary for Irish verbs has also been developed and made available on the Pota Focal website (Foclóir Briathra Gaeilge), providing additional insight into the labelling of verb particles.[13]

3 Categories of MWEs in the IUDT

The Irish Universal Dependency Treebank (IUDT) v2.5 (and earlier releases) contained many inconsistencies with respect to MWE annotations. This was mainly as a result of (i) automatic conversion from the IDT annotation scheme (ii) changes in the UD guidelines and (iii) a lack of broad coverage research into Irish MWEs. In this section we highlight the various UD labels for MWE categories (`compounds,` `fixed,` and `flat` (Nivre et al., 2020)) and how we have applied them to Irish in the v2.6 release. These categories are generally tailored towards those MWE types that demonstrate *syntactic idiosyncrasy*, i.e. they do not conform to the normal syntactic behaviour of the language. As such, many categories dealt with elsewhere (e.g. the verbal MWEs recognised in the PARSEME Shared Task) are not specifically addressed here. Our discovery/labelling approach involved flagging potential MWEs during treebank expansion annotation for the v2.6 release. Based on patterns we observed, we then actively searched for candidates for review in the v2.5 version. Compound labelling proved most challenging and therefore constitutes much of the discussion below.

[8] `https://universaldependencies.org`

[9] `https://universaldependencies.org/workgroups/mwe.html`

[10] `https://www.tearma.ie/`

[11] `https://www.focloir.ie/`

[12] `https://www.teanglann.ie/ga/`

[13] `http://www.potafocal.com/fbg/`

3.1 Compounds

It is difficult to define what is meant by the term ‘compound’, as, much like with MWEs themselves, different definitions abound (Marchand, 1960; Lieber and Štekauer, 2011; Fábregas and Scalise, 2012; Altakhaineh, 2016). Bauer (2001) offers the following definition: “We can now define a compound as a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the other(s) in other contexts, and which shows some phonological and/grammatical isolation from normal syntactic usage” (p. 695). However, unlike MWEs in general, a compound can either be written as a single combined word or a construction composed of multiple tokens.

Nivre et al. (2020) specifies that the compound relation should be used “for any kind of lexical compounds: noun compounds such as *phone book*, but also verb and adjective compounds, such as the serial verbs that occur in many languages, or a Japanese light verb construction such as *benkyō suru* (‘to study’).”

Compounds can pose problems for NLP analysis and their prevalence in written text along with their productive nature makes them an important consideration in computational processing (Ó Seaghdha, 2008; Nakov, 2013). However, compounds are often highly ambiguous and a large degree of “world knowledge” seems necessary to understand them, e.g. a cheese knife is a knife used to cut cheese, and not a knife made of cheese (Ó Séaghdha, 2007). We discuss the relevance of semantic compositionality further in Section 3.1.2.

3.1.1 Observation of variability across treebanks

The adoption of the `compound` relation and its subtypes varies across UD treebanks and this challenge is reflected in the extensive discussions on compounds on the UD MWE discussion forum.[14] In fact, there have been suggestions that only languages with a regular system of compounding should consider using the label. Based on our annotators’ linguistic backgrounds, we examined several treebanks and observed that the `compound` label is used relatively conservatively and inconsistently across UD treebanks, even within language groups:

French: The French documentation page notes that the compound label is seldom used.[15] This could be due to French syntax requiring prepositions to break up most seemingly compound nouns e.g. *tarte aux pommes* ‘apple tart’.[16] However, the `compound` relation is often used (instead of `fixed` or `goeswith`) for hyphenated words that have been split during tokenisation: e.g. *procès-verbal* ‘official report’, *outre-mer* ‘overseas’.[17] The label also appears to be used in rare cases in the French ParTUT treebank; *états membres* ‘member states’, *vote sanction* ‘protest vote’ and sometimes in cases that could be considered as nominal modifiers (`nmod`): *vendredi soir* ‘Friday evening’, in the French Spoken treebank. The leftmost noun is the head (right branching) in all cases. The `compound:prt` label is not used in any of the French treebanks.

Chinese: Chinese has several subtypes that reflect how compounding occurs naturally within the language, i.e. the `compound` label, used for noun-noun compounds such as 长途汽车‘long-haul coach’ and five subtypes which include; `compound:vv` used for verb-verb compounds, e.g. 找到‘to [try to] find’, as well as for verb-adjective compounds which are usually idiomatic, e.g. 他喊湿件恤衫‘He cried his shirt wet’; `compound:dir` verb-verb or verb-preposition compounds where one component is directional, e.g. 爬下来‘to climb down’; `compound:ext` used for the structural particle “得” in descriptive complements and complements of extent, e.g. 你门汉语说得很棒！‘Your Chinese is very good !’; and `compound:vo` which is used for verb-object compounds, e.g. 打电话‘to make a phone call’.

Spanish: Examples of the `compound` label use in the AnCora treebank include *por ciento* ‘per cent’; date strings: *el siglo XXI* ‘the 21st century’, *el año 2000* ‘the year 2000’, *el 6 de junio* ‘the 6th of June’;

[14] `https://universaldependencies.org/workgroups/mwe.html`

[15] https://universaldependencies.org/fr/dep/compound.html

[16] Example from French GSD treebank

[17] Examples from the UD-French-Spoken corpus.

light verb constructions: *tener en cuenta* 'to take into account', *dar derecho a* 'to authorize/to allow', *tener lugar* 'to take place', *poner fin (a)* 'to put an end (to)'; and verbal idiomatic constructions: *hacer hincapié* 'to emphasize the point'. The compound label is used in the GSD treebank for noun-noun compounds (e.g. *artista pop* 'pop star') and set phrases *hemisferio norte* 'northern hemisphere', *marca registrada* 'registered trademark' and *inmigrantes pos-apartheid* 'post-apartheid immigrants'. The label is also used for fixed adverbials *tal vez* 'perhaps', and for text in foreign languages (instead of flat:foreign) e.g. *nuevas mujeres (Xīn nˇüxìng, 1934)*. The use of the compound label therefore appears to be inconsistent across the sampled Spanish treebanks.

While the Spanish PUD treebank has no instances of the compound label it is the only Spanish treebank that uses the compound:prt label. It is applied to the reflexive pronoun *se* and all its forms, e.g. *Se ha recalcado* '(It) has been emphasised'.

English: The compound label is used in all the English Treebanks (GUM, GUMReddit, PUD, LinES, EWT, English-ESL) apart from the English-Pronouns Treebank. The rightmost noun is the head (left branching) in all cases. Noun phrases are marked as compounds in some treebanks (e.g. *wheel chair* (EWT) and *bank account* (PUD)). Adjectival compounds are labelled in some treebanks (e.g. *thought-provoking* (EWT), *Oscar-winning* (PUD) and *Dutch speaking* (GUM), *self-driven* (EWT)). The compound label is also used for numbers in some treebanks (e.g. *two hundred* (EWT), *3 million* (PUD)). Some named entities are also labelled as compound (e.g. *United States* (PUD) and *Auckland Castle* (GUM)). The compound:prt label is used for some phrasal verbs in PUD & EWT (e.g. *sign up, point out*).

3.1.2 Labelling compounds in the Irish UD treebank

Compounds in Irish are briefly dealt with by the Christian Brothers (1999). According to their analysis (p.277), Irish compounds are formed in two ways:

(i) when a prefix or suffix is added to a word to form a single-token compound (e.g *idirdhealaigh* 'to differentiate', *idir-* 'inter' + *dealaigh* 'to detach/to separate').

(ii) two words appear together to form a compound (adjectival, verbal and nominal) – which is of most interest to our work here. These can occur as (a) a single token (e.g. *cúl* 'back' + *caint* 'talk' → *cúlchaint* 'gossip') or (b) multiword compounds *mac léinn* ('student', lit. son of learning), *mac tíre* ('wolf', lit. son of land). We are only concerned with annotation of the second type (b) here. This type of compound is less frequent in Irish than the single-token type and would therefore suggest a conservative use of the label in the treebank.

In the Irish UD treebank, we apply two UD compound labels: the standard dependency relation label compound and the subtype label compound:prt. It should be noted that due to Irish word order, the head noun of a noun phrase is usually the first noun, hence the compounding attachment is right-branching. The following reports on the criteria used for determining compound constructions in the Irish UD treebank, while aligning as closely as possible to the UD documentation and conventions.

Compound label The compound label is used exclusively for labelling noun-noun compound constructions in the IUDT. Applying the compound label proved challenging in the Irish data. Initial annotation discussions revealed different interpretations as to what would be deemed a compound noun in Irish. Often a disagreement arose with respect to two nouns appearing together that some believed to be nominal modifiers (nmod e.g. *fonn díoltais* 'revengefulness') as opposed to compounds.

As a first step to identifying Irish compound nouns, a number of tests, inspired by the PARSEME guidelines,[18] were used as a preliminary attempt at determining whether an Irish nominal multiword unit should be marked as compound or nmod. However, while helpful to some degree, it transpired that no test on its own was sufficient to decide this.

Test 1: The absence of a definite article We theorised that the absence of a definite article between nouns could indicate a nominal compound. There is no indefinite article in Irish. Therefore, the article

[18] https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Criteres

is absent in noun-noun compound constructions where the dependent noun is indefinite (e.g. *deireadh seachtaine* 'weekend' (lit. end of week)), which we label as `compound`. The same construction with a definite noun dependent would usually be labelled as a nominal modifier `nmod` (*deireadh **na** seachtaine*: 'end of **the** week'). Other examples of this phenomenon include *caitheamh aimsire* 'pastime' vs *caitheamh **na** haimsire* 'passing of **the** time' and *tiarna talún* ('landlord') vs *tiarna **na** talún* ('lord of **the** land').

This test cannot be reliable in all cases where the definite article is absent, however, as the article is also dropped in other instances (1) the possessive genitive case; *pá iriseora* 'a journalist's pay' (lit. pay of a journalist); (2) when a noun is followed by another qualified noun *pá iriseoir **an** nuachtáin* 'the newspaper's journalist's pay' (lit. pay of **the** journalist of **the** newspaper). We would consider both (1) and (2) to be nominal modifiers `nmod`.

Furthermore, there are other exceptions that affect the reliability of this test and these are usually idiomatic or institutionalised phrases; *Mí na meala* 'honeymoon' (lit. month of the honey); *cothrom na Féinne* 'fair treatment' (lit. balance of the warriors).

Test 2: Presence of a cranberry word If a noun-noun construction contains a *cranberry* word (i.e. a word that does not occur outside of that specific construction) (Aronoff, 1976), it is a strong indicator that the construction should be labelled as `compound`. For example, *déag* 'teen' only occurs as part of a numeral phrase (*cúig déag* 'fifteen' (lit. 'five teen')). While cranberry words can occur in other types of phrases (e.g. adverbial, *go deo* 'forever', which is annotated as `fixed`, see Section 3.2), *déag* remains the only cranberry word identified in a nominal construction in the Irish UD treebank and accounts for a small percentage of compounds in our data.[19]

Test 3: Determining if the meaning of either noun is sufficiently changed Nakov (2013) notes non-compositionality as semantic criteria for compounds and that this criterion asks that compounds be at least partially non-compositional. However, he also notes that compositionality is a matter of degree. In a compound construction, the meaning of the whole must be significantly different from the meaning of the individual tokens in the noun phrase. Compositionality, therefore, appeared to be a good basis for assessing compound candidates and finding agreement amongst annotators: `Fully compositional`: the meaning as a whole can be easily interpreted from the meaning of each of the parts of the multiword construction (e.g. *turas scoile* 'school trip'); `Semi-compositional`: the meaning of the whole expression can be partially understood from the meaning of the individual parts e.g. *lucht leanúna* 'followers' (lit. group of following) or *feadóg stáin* 'tin whistle' (that can be made of wood); `Non-compositional`: the meaning of the unit or expression as a whole is not discernible from the meaning of the individual parts. e.g. *mac tíre* 'wolf' (lit. son of land).

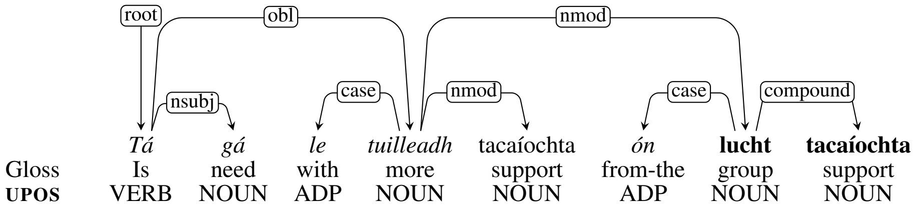
We carried out an anonymous poll to categorise 30 contentious nominal multiword units in terms of compositionality. Six participants with varying levels of fluency and knowledge of Irish syntax reviewed the candidates, and selected one of the three compositionality measures for each one. Based on agreement levels, it was established that this compositional categorisation approach was useful for compound categorisation. Table 1 shows that agreement on whether constructions were fully-compositional compounds rather than semi- or non-compositional compounds was easier to achieve than differentiating between semi- and non-compositional compounds.[20] This is significant for differentiating between `compound` and `nmod`. The overall compositionality scores for each candidate were averaged across the 6 annotators. If the average score for fully-compositional was higher than 0.5, the `nmod` label was applied. Otherwise, (i.e. the average score for either a semi-compositional or non-compositional label was greater than or equal to 0.5), the `compound` label was applied. Figure 1 shows the dependency annotation of an Irish semi-compositional compound.

[19] There are 20 occurrences of *déag* 'teen' in v2.6 of the IUDT.

[20] A possible reason for disagreement in cases with a clear majority is a potential confusion on the part of a survey participant of the terms fully and non-compositional, e.g. *mac tíre* is clearly idiomatic and should be regarded as non-compositional.

Compound candidate	Literal meaning	Translation	F.C.	S.C.	N.C.
deireadh seachtaine	‘end of week’	weekend	2	3	1
mí na meala	‘month of the honey’	honeymoon	1	1	4
mac tíre	‘son of land’	wolf	1		5
mac léinn	‘son of learning’	student	1	1	4
lucht féachana	‘people of watching’	audience	3	3	
caitheamh aimsire	‘spending of time’	past time	2	4	
cothrom na Féinne	‘balance of the Fianna’	fair treatment		3	3
tús áite	‘start place’	priority	1	5	
feadóg stáin	‘whistle of tin’	tin whistle	1	5	
foinse saibhris	‘source of wealth’	source of wealth	5		1
fonn díoltais	‘vengeful mood’	vengeful mood	4	1	1
cumas an pháiste	‘the ability of the child’	the ability of the child	5		1

Table 1: Poll results on 6 annotators’ opinions on compositionality of some controversial compound candidates. F.C. Fully Compositional; S.C Semi-compositional; N.C. Non-compositional
Blue text indicates categorisation as `compound`.



‘There is a need for more support from the supporters’

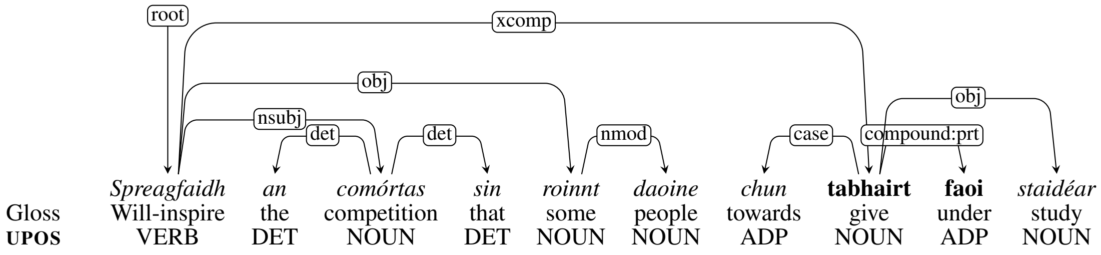
Figure 1: Dependency annotation of semi-compositional compound

Compound:prt label We apply the `compound:prt` label to verb particles as per the UD guidelines. [21] Verb particle constructions consist of a verb and a dependent particle, where the particle significantly changes the meaning of the verb. In Irish, this particle is usually a directional adverb, although certain prepositions can also function as particles (e.g. *faoi* ‘under/beneath’, *as* ‘off/out’). As both the particle and the verb are necessary for understanding of the construction as a whole, we consider these constructions as MWEs. Table 2 provides examples of Irish verb particles and Figure 2 shows an example in terms of dependency tree annotation.

	Gloss	Literal meaning
tabhair suas	‘to give up’	(give + up)
tabhair faoi	‘to undertake’	(give + under)
éirigh as	‘to retire’	(rise + from)
éirigh amach	‘to revolt’	(rise + out)
leag amach	‘to outline’	(lay + out)
leag síos	‘to lay out’	(lay + down)
bain amach	‘to get/to reach’	(extract + out)
dul as	‘escape’	(go + from)

Table 2: Examples of Irish verb particles labelled as `compound:prt`

[21] `https://universaldependencies.org/docs/en/dep/compound-prt.html`

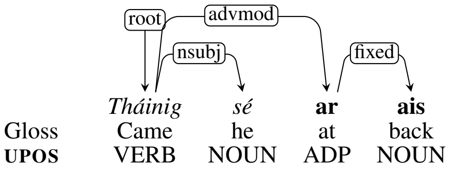

root
xcomp
obj
nsubj
det
det
nmod
case
compound:prt
obj
Spreagfaidh an comórtas sin roinnt daoine chun tabhairt faoi staidéar
Gloss Will-inspire the competition that some people towards give under study
UPOS VERB DET NOUN DET NOUN NOUN ADP NOUN ADP NOUN

‘The competition will inspire some people to undertake studies’

Figure 2: Dependency annotation of verb particle

3.2 Fixed

Fixed expressions are a class of fully lexicalised immutable expressions that are generally non-compositional in nature (Sag et al., 2002). Flexibility is a widely used characteristic for determining fixed expressions, which refers to the potential for components of the expression to inflect for gender, number, etc. Examples of fixed expressions in English include set phrases (*of course*), function words (*as if*) and/or short adverbials (*at least*). Within the UD annotation scheme, the `fixed`[22] dependency label is used for “certain fixed grammaticized expressions that behave like function words or short adverbials” and it is assumed that “these expressions do not have any internal syntactic structure (except from a historical perspective)”.[23] The dependents within the fixed construction each attach to the leftmost token (the head), as per Figure 3.



‘He came back’

Figure 3: Dependency annotation of fixed multiword adverbial modifier

With regards to the application of the `fixed` label in the Irish treebank, it can be said that Irish fixed MWEs have the following characteristics:

Fixed grammaticized expressions Such fixed expressions perform a specific grammatical role (e.g. adverbial modifiers such as *go deo* ‘forever’, *mar sin* ‘so’, *ar ais* ‘back’; or nominal modifiers like *a thuilleadh* ‘more’ and *go leor* ‘enough’).

Neither semi-fixed, flexible or productive It is generally considered that the flexibility of the tokens within the MWE determine whether or not they are fixed. For example, we currently treat *chomh maith* ‘as well’ as a fixed adverbial. The preposition *le* ‘with’ often follows this expression. However, *le* may inflect for gender and person (as a pronominal preposition, e.g. *liom* ‘with me’) and given that such multi-token units are not currently split, it is therefore too flexible to be considered part of the fixed expression. Likewise, no tokens can be inserted within the fixed expression (e.g. **faoi an bhun*).

No discernible internal syntactic structure In spite of the term used, we annotate compound prepositions as `fixed` and not as `compound`, given that they are often a combination of preposition and noun, yet their joint functional role is that of preposition (e.g. *in aice* ‘near’; *tar éis* ‘after’; *i gcoinne* ‘against’;

[22] Formerly the `mwe` label in UD v1 guidelines `https://universaldependencies.org/docsv1/`

[23] `https://universaldependencies.org/u/dep/fixed.html`

de réir 'according to'). They also select for a prepositional object (e.g. *in aice na tine* 'beside the fire'; *de réir na comhairle sin* 'according to that advice').[24]

MWEs labelled as `fixed` in the Irish treebank are outlined in Table 3 and grouped according to their syntactic role.

Fixed MWE	**Gloss**	**Fixed MWE**	**Gloss**
Adverbial Modifier		**Prepositional (case)**	
go deo	'forever'	*in aice*	'beside'
mar sin	'so'	*tar éis*	'after'
ar ais	'back'	*ar fud*	'throughout/ all over'
chomh maith	'as well'	*le haghaidh*	'for'
fad is	'as long as'	*le linn*	'during'
a mhéid	'to the extent that'	*faoi bhun*	'below'
Determiners		*go dtí*	'towards/to'
seo caite	'last'	*de bharr*	'because of/ due to'
seo chugainn	'next'	*i rith*	'during'
Nominal Modifier		*i gceann*	'at the end of'
ar bith	'at all'	**Open Complement (xcomp:pred)**	
Subject/Object		*in ann*	'able to'
a thuilleadh†	'more'	**Subordinating Conjunction**	
go leor†	'a lot'	*le go*	'so that'

Table 3: Types of Irish MWEs labelled with the UD `fixed` dependency label. † indicates that these MWES can also function as adverbials.

3.3 Flat

According to the UD guidelines, the `flat` dependency relation is used for exocentric (or headless) semi-fixed MWEs, such as personal names and dates strings. The assumption is that these expressions do not have any internal syntactic structure and that the structural annotation is in principle arbitrary. Therefore, flat MWEs are annotated with a flat structure, where all subsequent tokens in the expression are attached to the first token using the `flat` label.

The `flat` relation and its subtypes `flat:foreign` and `flat:name` (see below) are all therefore used for headless semi-fixed MWEs. MWEs that come under this category vary widely across languages but a "regular compositional syntactic structure" is assumed when using the flat label.[25] Examples from the Irish UD treebank include days of the week (*Dé Luain*, 'Monday') and dates (*Deireadh Fómhair* 'October'; *(I) mí Iúil 1995* '(In) July 1995'; *(roimh) 1 Feabhra 1997* '(before) February 1 1997'). The internal components of these MWEs are attached to the left-most token of the noun phrase using the flat label.

The UD guidelines state that "For organization names with clear syntactic modification structure, the dependencies should also reflect the syntactic modification structure using regular syntactic relations, as in: 'Lord of the Rings"'. However, it should be noted that we diverged from this temporarily in an exercise in capturing Named Entity (NE) information during the MWE review. We currently use the `flat` relation in v2.6 for named entities and proper noun strings regardless of whether or not their internal syntactic structure is discernible. Some examples include: organisations such as *Choiste Turasóireachta na Gaillimhe* 'Galway Tourism Board' and *Roinn na Gaeltachta* 'Department of the Gaeltacht'; titles such as *Ard-Cheannasaí* 'High Commander' and *Mharascal Machaire* 'Field Marshal', titles of published works such as *Leatrom na Cinniúna* 'The Injustice of Destiny', placenames such as *Baile Átha an*

[24] It should be noted that prior to conversion to UD style, fixed expressions in the IDT were treated as one token joined together by an underscore in the data (e.g. *in_aice)* as per the output of the standard Finite-State Irish POS-tagger (Dhonnchadha, 2002).

[25] `https://universaldependencies.org/u/dep/flat.html`

Rí 'Athenry' and other named entities such as *Bunscoil Mhic Reachtain* 'McCracken Primary School'.[26]

flat:name The use of this label is reserved for personal name strings whereby the first nominal token is labelled as the head and its subsequent tokens in the string are annotated as `flat:name` (e.g. *Pádraig Mac Piarais* 'Patrick Pearse'). The various name particles used in Irish (*Ó, Ua, Mac, Mag, Uí, Ní, Mhic, Nic, Nig*) are also assigned the `flat:name` label, as are professional titles (e.g. *An tUasal* 'Mr', *Dochtúir* 'Doctor', *T.D. (Teachta Dála)* 'Member of the Irish Parliament')). Therefore, in the personal name *Liam Ó Briain*, both the particle *Ó* and the proper noun *Briain* are both attached to *Liam* with the dependency relation `flat:name`.

flat:foreign This label replaces the `foreign` label that was used in v1 of the UD guidelines.[27] It is used for words in other languages that appear in a linear sequence, including cases where foreign text is incorporated into a sentence, e.g. *go raibh sé cut off with a shilling* 'that he was cut off with a shilling'. The treatment of proper noun strings such as personal names and titles in other languages that occur in the data (e.g. Monsieur Dupont, 'Nor Meekly Serve My Time') raised a question amongst annotators (i.e. `flat` vs `flat:foreign`). The approach taken was that personal names, regardless of their origin (e.g. Bertie Ahern, *an tUasal* Durkan, Mr Mulligan, Robert de Niro, *an tUas* Morten Kjaerum) are to be labelled with the `flat:name` relation (see above). Titles in a foreign language (e.g. 'The Pope's Green Island', 'Entering Jerusalem', 'Tristan Und Isolde') should be treated as `flat:foreign`. Multiword expressions in other languages are also labelled as `flat:foreign`, e.g. *vice versa*.

Another interesting question relates to the POS-tagging of foreign words in the treebank. In previous IUDT releases, English tokens were POS-tagged according to English morpho-syntax (e.g. NOUN, PROPN, etc). However, according to current UD guidelines, the X tag should be used for foreign words. Nevertheless, in future releases it is hoped that English tokens will be re-annotated with their appropriate POS tag to allow for more concise code-switching studies, as per recent recommendations by Sanguinetti et al. (2020).[28]

4 Conclusion

In this article we have reported on a review and update of MWE annotations in the Irish UD Treebank for the v2.6 UD release. We have provided our analysis and motivations for applying the `compound`, `fixed` and `flat` labels to Irish MWEs, and discussed the various challenges involved therein. In the v2.6 treebank of size 64,745 tokens, the `compound` label was applied 160 times (141 `compound`, 19 `compound:prt`), `fixed` was applied 950 times, and `flat` was applied 2252 times (1399 `flat`, 695 `flat:name`, 150 `flat:foreign`).

While our approach is mostly in line with the UD annotation guidelines, we note that our use of the `flat` label is too broad as it also incorporates Named Entities (NE) in general. The opportunity for manual review of the treebank data allowed for previously unknown NE data to be captured easily. In the future, we want to remove the `flat` label in these cases and capture NE information in the MISC column instead. Finally, a note on inflected prepositions (see Section 3.2). Currently we do not split pronominal prepositions into ADP + PRON. If however, future versions of the treebank undergo changes with respect to splitting multi-token units (e.g. *leis* 'with it' → *le* + *é*), the uninflected preposition token could be considered part of a fixed expression (e.g. *chomh maith le*).

Acknowledgements

This work is funded by the Irish Government Department of Culture, Heritage and the Gaeltacht under the GaelTech Project, and also supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) at Dublin City University. We would like to thank Jennifer Foster, Lauren Cassidy, Mícheál John Ó Meachair and Carla Parra Escartín for their input to our linguistic discussions.

[26] Plans to review this approach are underway, with the consideration of using the MISC column instead to capture NE information, as per the English-GUM treebank.

[27] `https://universaldependencies.org/docsv1/u/dep/foreign.html`

[28] In their work on treebanks for user-generated content, they propose appropriate POS-tagging of foreign text along with an indication (LangID=EN in the MISC column) that code-switching has taken place, when the language is known to annotators.

References

Abdel Rahman Altakhaineh. 2016. What is a compound? the main criteria for compoundhood. *Explorations in English Language and Linguistics*, 4, 10.

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. MIT Press, Cambridge, Massachusetts and London, England.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. *Handbook of Natural Language Processing, Second Edition*, pages 267–292, 01.

Laurie Bauer. 2001. *Compounding*. Language Typology and Language Universals, volume 1. Walter de Gruyter, Berlin.

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, 01.

Eduard Bejček and Pavel Straňák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44:7–21, 04.

Maria Bloch-Trojnar. 2009. On the Nominal Status of VNs in Light Verb Constructions in Modern Irish. In *PASE Papers 2008. Vol. 1: Studies in Language and Methodology of Teaching Foreign Languages*, page 25–33, Wrocław: Oficyna Wydawnicza ATUT.

Marie Candito and Mathieu Constant. 2014. Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing. In *ACL 14 - The 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, United States, June. ACL.

Christian-Brothers. 1999. *Graiméar Gaeilge na mBráithre Críostaí*. An Gúm, Baile Átha Cliath.

Matthieu Constant and Joakim Nivre. 2016. A Transition-Based System for Joint Lexical and Syntactic Analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 161–171, 01.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 204–212, Jeju Island, Korea, July. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892, December.

Matthieu Constant, Gülşen Eryiğit, Carlos Ramisch, Mike Rosner, and Gerold Schneider, 2019. *Statistical MWE-aware parsing*, pages 147–182. Berlin: Language Science Press, 01.

Elaine Uí Dhonnchadha. 2002. Two-level Finite-State Morphology for Irish. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 2299–2306, Gran Canaria, Spain. European Language Resources Association (ELRA).

Elaine Uí Dhonnchadha. 2009. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. Ph.D. thesis, Dublin City University.

Gülşen Eryiğit, Kübra Adali, Dilara Torunoğlu-Selamet, Umut Sulubacak, and Tuğba Pamay. 2015. Annotation and extraction of multiword expressions in Turkish treebanks. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 70–76, Denver, Colorado, June. Association for Computational Linguistics.

Antonio Fábregas and Sergio Scalise. 2012. *Morphology: From Data to Theories*. Edinburgh Advanced Textbooks in Linguistics. Edinburgh University Press.

Christiane Fellbaum. 1998. A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32(2/3):209–220.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, 39(1):195–227.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. Linguistic Inquiry monographs volume 28. MIT Press.

Sylvain Kahane, Martine Courtin, and Kim Gerdes. 2018. Multi-word annotation in syntactic treebanks: Propositions for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 181–189, Prague, Czech Republic, 01.

Rochelle Lieber and Pavol Štekauer. 2011. Introduction: Status and definition of compounding. In *The Oxford Handbook of Compounding*, pages 3–18, Oxford: Oxford University Press.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Teresa Lynn and Jennifer Foster. 2016. Universal Dependencies for Irish. In *Proceedings of the Second Celtic Language Technology Workshop*, Paris, France.

Teresa Lynn, Jennifer Foster, and Mark Dras. 2017. Morphological features of the Irish Universal Dependency Treebank. In *TLT 2017 : Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*, volume 1779, pages 111–122, Bloomington, U.S.

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

Hans Marchand. 1960. *The Categories and Types of Present-day English Word-formation: A Synchronic-diachronic Approach*. Wiesbaden: Otto Harrassowitz.

Héctor Martínez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98, 09.

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3), 07.

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

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, 01.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Diarmuid Ó Séaghdha. 2007. Annotating and Learning Compound Noun Semantics. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 73–78, Prague, Czech Republic, June. Association for Computational Linguistics.

Diarmuid Ó Seaghdha. 2008. *Learning Compound Noun Semantics*. Ph.D. thesis, University of Cambridge.

Carla Parra Escartín, Almudena Nevado, and Eoghan Martínez. 2018. Spanish multiword expressions: Looking for a taxonomy. In *Multiword expressions: Insights from a multi-lingual perspective*, pages 271–323. Berlin: Language Science Press, 05.

Behrang Qasemizadeh. 2014. Annotation of Multiword Expressions in the Farsi Section of the Universal Dependencies Project. Second PARSEME General Meeting Posters, March.

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. ACL.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.

Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. 2016. MWEs in Treebanks: From Survey to Guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2323–2330, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Victoria Rosén, Gyri Smørdal Losnegaard, De Smedt Koenraad, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Manfred Sailer, and Mitetelu Verginica. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories*, page 179–193, Warsaw.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. pages 1–15, 02.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoglu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference,LREC*, pages 5240–5250, Marseille, France, May. European Language Resources Association.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November.

Agata Savary, Carlos Ramisch, Silvio 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. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain, April. Association for Computational Linguistics.

Gerold Schneider. 2014. Improving PP attachment in a hybrid dependency parser using semantic, distributional, and lexical resources. In *Second PARSEME Meeting*, Athens, Greece.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 2371–2378, Portorož, Slovenia, 05. European Language Resources Association (ELRA).

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction.* Text, Speech and Language Technology volume 44. Springer Netherlands.

Kresimir Sojat and Matea Filko. 2016. Verbal Multiword Expressions in Croatian. In *Proceedings of the Second International Conference Computational Linguistics in Bulgaria (CLIB 2016)*, pages 78–85, 09.

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

Umut Sulubacak and GülşenT Eryiğit. 2018. Implementing Universal dependency, morphology, and multiword expression annotation standards for Turkish language processing. *Turkish Journal of Electrical Engineering and Computer Sciences*, 26:1662–1672, 05.

Veronika Vincze, János Zsibrita, and T. IstvánNagy. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proceedings of the International Joint Conference on Natural Language Processing*, page 207–215, Nagoya, Japan, 10.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2019. Ilfhocail: A lexicon of Irish MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 162–168, Florence, Italy, August. Association for Computational Linguistics.

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