Modeling Syntactic Dependency Relationships in Wikidata Lexicographical Data

Mahir Morshed

University of Illinois at Urbana-Champaign, Urbana, IL 61801
mmorshe2@illinois.edu

Abstract. We present a scheme with which a lexeme on Wikidata consisting of multiple parts may be annotated to denote syntactic dependencies among its parts. The scheme is sufficiently general to accommodate many dependency grammar frameworks and can take advantage of Wikidata lexemes' structure to reduce redundancy in representation while still being flexible enough for further qualification. While we note some challenges in adjustments to the scheme for particular phenomena, we contend that adopting this scheme will aid syntactic parsing efforts in other general domains as well as text generation systems for the Abstract Wikipedia project.

Keywords: lexicographical data · syntax · dependency grammar

1 Introduction

The projects under the Wikimedia Foundation's umbrella have frequently been used for various natural language processing tasks, including disambiguating word senses [2], recognizing named entities [7], and for low-resourced languages potentially many others [9]. Some efforts at syntactic annotation of text from these projects also exist [3, 6], but these typically infer grammatical information from the text ingested based on systems with some prior acquired syntactic reasoning, rather than retrieve this information directly from textual elements.

The under-construction Abstract Wikipedia project [11] has as a goal the ability to generate text in any natural language from a representation constructed purely of abstract concepts, these concepts transformed via language-specific renderers into some textual representation. The building blocks of this text are planned to be Wikidata *lexemes*–objects corresponding to units of linguistic meaning (primarily words, but also expressions with multiple parts such as compound words, idioms, and proverbs). These lexemes are similarly structured to Wikidata items, but they are modeled in a separate namespace, have special fields for lemmata, language, and lexical category, and have separate substructures for different meanings (*senses*) and inflectional realizations (*forms*).

For these building blocks to be useful, some mappings from concepts to lexemes must first exist, which presently consist of synonym and translation linkages between senses and correspondences between Wikidata items and senses. A concept that in one language is representable with one word may need multiple words in another, however; depending on the sort of multi-part expression used, the syntactic information needed for adjustment of that expression in different contexts may differ. The English verb 'evade' has a correspondence with the South American Spanish phrase 'hacer el quite' [1], for example; 'el quite' appears to play a role similar to an object of the verb 'hacer' and could be marked and adjusted as such within a sentence. Not only may other equivalents between languages behave even more differently, but the composition of lexemes to represent more complex concepts only yields non-decreasing potential syntactic differences.

Databases of multi-component expressions have previously been developed for individual languages [4, 5, 8], most primarily focusing on annotating phrase structure constituency relations, but some also doing so with dependency grammar relations [10]. Although the multilingual structured nature of Wikidata's lexicographical data makes it attractive as a place to aggregate such databases, the structure of Wikidata lexemes and their statements makes annotating constituencies directly difficult: the overhead for storing each of a phrase structure tree's intermediate levels, whether as separate lexeme statements or as entirely separate objects, and compared to storing dependency information, may be much greater than necessary for the Abstract Wikipedia project and other language generation applications.

We thus propose here a compact representation of syntactic dependencies within the structures of Wikidata lexicographical data, generally applicable to different flavors of dependency grammar, but here demonstrated with respect to Universal Dependencies (UD). We contend that the marking up of this information is useful even for modeling structures of multi-part elements that may be regarded in some languages as words, and that it permits lexemes with syntax represented this way to form parts of other lexemes which, as single units, take part in other dependency relations. We recognize too that modifications to handle special syntactic cases may not necessarily be immediately acceptable to those annotating relevant lexemes. We nevertheless believe that the greater portion of what may be annotated of multi-part lexemes with this representation will not only make those lexemes usable in the syntactic parsing of other texts, but will also considerably ease the generation of text through the manipulation of underlying dependency graphs.

2 Implementation

What follows is an outline of the proposed dependency representation within Wikidata lexemes. RDF predicates for Wikidata properties and qualifiers[1] are

[1] https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format (section "Full list of prefixes") lists the RDF prefixes used herein.

provided in monospace, as are Wikidata items, lexemes, and their forms and senses, which are left unprefixed.

	what (L333986)	go (L3006)	around (L333609)	come (L3210)	around (L333609)
'series ordinal'	1	2	3	4	5
'object form'	what (L333986-F1)	goes (L3006-F2)	around (L333609-F1)	comes (L3210-F2)	around (L333609-F1)
'object sense'	that which (L333986-S1)	to move away (L3006-S2)	at varied places (L333609-S1)	to approach (L3210-S1)	at varied places (L333609-S1)
'head position'	2	4	2	0	4
'head relationship'	relativizer (Q56870226)	subject clause (Q19708532)	location adverbial (Q12724480)	root (Q1757074)	location adverbial (Q12724480)
UD equivalent of 'head relationship'	mark	csubj	obl:lmod	root	obl:lmod

Fig. 1. Breakdown of the set of five qualifiers on the 'combines' statements on the lexeme for the proverb 'what goes around comes around' (L345525); the first three are standard for 'combines' statements, and the newly created fourth and fifth mark syntactic dependencies.

2.1 Lexeme components

Each component of the surface form of a multi-part lexeme is represented by a use of the 'combines' property (p:P5238) linking to a lexeme for said component. The component's representation in the lexeme is specified with a qualifier noting the form which is the object of the 'combines' statement ('object form', pq:P5548) on each statement, and the position of each part using a 'series ordinal' qualifier (pq:P1545). For completeness, the statement may also note the sense of the component being used with 'object sense' (pq:P5980); language-specific machinery for text generation may find utility in different treatment of a constituent based on the meaning it expresses, although we do not further consider uses of that qualifier here.

An example of these qualifiers in action is shown in the first three rows of Figure 1 for the proverb "what goes around comes around". Note that 'series ordinal' values need not be numeric; a separate ordering to handle infixes, circumfixes, and other non-sequential phenomena may well be desired for some languages.

Many dependency treebanks optionally store information about the part of speech of a word token and the grammatical features it bears in the context in which it appears. In a Wikidata lexeme, the former of these is a top-level feature (via `wikibase:lexicalCategory`), while the latter of these reside on the individual forms of lexemes (via `wikibase:grammaticalFeature`). As a result this information for the components of a multi-part lexeme generally does not need to be reproduced on the multi-part lexeme itself; they may be programmatically retrieved from the parts themselves based on the main 'combines' values and 'object form' qualifiers respectively.

2.2 Syntactic annotation

A dependency relationship may be thought of as a directed edge between a dependent and a head, so that specifying both ends of the relationship and the type of relationship suffices to define the dependency, and so that the resulting set of dependencies for a multi-component lexeme resembles a tree. To define dependency relationships between parts of a lexeme within Wikidata's existing structure, a set of qualifier properties are instead applied to the 'combines' statement for a relationship's dependent.

The first of these qualifiers, 'head position' (`pq:P9764`), indicates the 'series ordinal' of the head of a dependency relationship. The second of these, 'head relationship' (`pq:P9763`), indicates the type of said relationship. Both of these qualifiers, having been created in late July 2021, are as yet little used beyond additions of these by the author on existing lexemes, and user scripts to better facilitate their addition to new lexemes on Wikidata are yet to be written. Documentation of appropriate 'head relationship' values is slowly being developed, however[2].

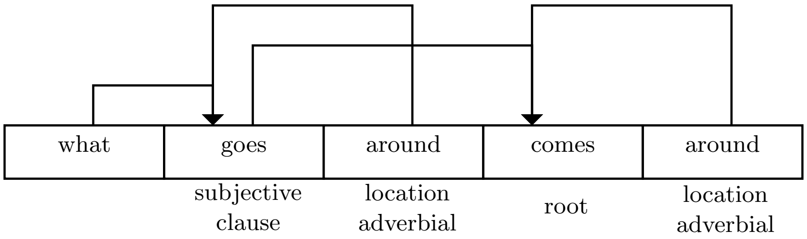

Fig. 2. Diagram of the relationships between components of 'what goes around comes around' induced by the statements and 'head position' qualifiers noted in Figure 1, with 'head relationship' values for each arc provided below each arc's origin. All arcs point from a relationship's dependent to the relationship's head.

The relationships in the proverb 'what goes around comes around' are defined in the fourth and fifth rows of Figure 1, with UD equivalents below them. A potential diagram generated with 'series ordinal' and the two new qualifiers is

[2] With respect to UD, see https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Universal_Dependencies .

shown in Figure 2. The link from “what” to “goes” is defined by the edge from ‘series ordinal’ value ‘1’ to ‘series ordinal’ value ‘2’, where this latter value is specified via the ‘head position’ value on “what”. Most other relationships are marked up similarly. As a convention, the ‘head position’ of the root part is ‘0’ and the ‘head relationship’ is the item for ‘root’.

The two qualifiers, as two separate properties defining parts of the same relationship, were proposed because Wikidata statements and qualifiers cannot have statements or qualifiers themselves as values, much less annotate connections between them, without major modifications to Wikidata’s Wikibase software and its Query Service. While such connections might also be stored in an entirely separate Wikibase instance, in the absence of an implicit federation system between Wikibase triple stores, querying such connections becomes more costly for end users and downstream applications than necessary.

3 Potential challenges

With this dependency representation, a large number of syntactic phenomena in multi-part lexemes can be faithfully modeled. There nevertheless remain some situations that an application of this representation by itself would not satisfactorily handle, and for which the introduction of certain changes might be controversial. We outline some of these challenging aspects here.

3.1 Elided components

In some parallel constructions, one may decide to omit in later parts of a phrase portions common to earlier parts of the same phrase. The sentence “I wrote the book, he wrote the story, and she wrote the poem” may be shortened without loss of understanding by omitting the latter two occurrences of ‘wrote’. The resulting subgraphs of the latter phrases may be marked differently as well; in UD the special orphan relation would link ‘he’ and ‘the story’, as well as ‘she’ and ‘the poem’, in the shortened version of the sentence.

	I	write	the	book	and	he	*no value*	the	story
‘series ordinal’	1	2	3	4	5	6	7	8	9
‘head position’	2	0	4	2	7	7	2	9	7
‘head relationship’	subject	root	determiner	direct object	conjunction	subject	conjunct	determiner	direct object
UD equivalent of ‘head relationship’	nsubj	root	det	obj	cc	nsubj	conj	det	obj

Fig. 3. Breakdown of some qualifiers on the ‘combines’ statements for the hypothetical lexeme ‘I wrote the book and he the story’, where IDs for lexemes, forms, and items have been omitted for brevity.

UD alternatively defines, in its specifications of entirely optional ‘enhanced dependencies’, the concept of ‘orphan nodes’ which represent elided words and can take part in those dependency relationships originally substituted with orphan. To mimic this concept, elided words might be similarly indicated by setting the value of the 'combines' statement to the special Wikibase value "no value" and otherwise marking up relationships with respect to that elided word. (In the interest of preserving some contextual information, the 'object form' qualifier might still refer to the form the elided part would take on were it still present in the lexeme.) An example of an elided word's use is shown in Figure 3.

The insertion of extra 'nonexistent' components to a multi-part lexeme may appear to some to be a repurposing of the 'combines' property, given that these components do not appear in the surface form of that lexeme and that any counts of components of such a lexeme may appear inflated. We might alternately contend that this is merely a syntactic analogue of when grammatical features added to a lexeme form may not necessarily change the form despite those features' importance (as, for example, when further inflecting a Hindustani adjective that is already inflected for female gender), and that in counting components filtering out "no value" statements is a small addition to a query.

3.2 Cross-clausal relationships

	She	and	I	visited	Japan	and	Korea
'series ordinal'	1	2	3	4	5	6	7
'head position'	4	3	1	0	4	7	5
'head relationship'	subject	conjunction	conjunct	root	object	conjunction	conjunct
UD equivalent of 'head relationship'	nsubj	cc	conj	root	obj	cc	conj
'object has role'			subject				object

Fig. 4. Breakdown of some qualifiers on the 'combines' statements for the hypothetical lexeme 'She and I visited Japan and Korea', where IDs for lexemes, forms, and items have been omitted for brevity.

A clause may contain multiple instances of the same type of element, such as two subjects or two objects, to which other components in the clause apply equally. "She and I visited Japan and Korea" contains two subjects, each of which applies to two objects, and "They washed and combed the dog" contains two predicates. The graph of the first phrase may directly connect 'she' to the predicate and connect 'I' to 'she' (the approach taken by UD), and the graph of the second phrase may alternately group the actions 'washed' and 'combed' into an umbrella predicate to which 'they' and 'the dog' directly attach. Either of these approaches, however, adds distance between components that we might regard as being close syntactically.

UD's 'enhanced dependencies' also allow multiple relationships to share the same dependent, so that in the first example the word 'I' points both to 'she' (as a conjunct) and ‘visited’ (as a subject), turning the resulting syntactic tree into a general directed (possibly cyclic) graph. Since multiple ‘head position’ and ‘head relationship’ qualifiers on a single ‘combines’ statement cannot be separated to refer to different dependency relationships, one possible solution is to mark out the semantic roles more explicitly, and in some cases redundantly, using the ‘object has role’ (`pq:P3831`) or ‘has quality’ (`pq:P1552`) qualifiers on the ‘combines’ statements of conjuncts. An example using the first phrase and the ‘object has role’ qualifier is shown in Figure 4.

The extra marking of syntactic roles may appear to some to simply duplicate information, especially since a conjunct of a particular part very frequently has the same role as that part. We might instead say that the explicit marking of the relationships expressed by conjuncts allows them to be queried more readily, so that traversing conjunct paths becomes unnecessary.

3.3 Echo words and reduplication

In many South Asian languages, it is a frequently productive process for a word and a nonce word rhyming with it, when taken together, to refer to something related to said word. In Bengali the noun “ranna” is the act of cooking, while “rannabanna” refers to ‘cooking and related activities’; in Hindi “samna” is ‘to encounter’, while “amna samna” is the act of encountering. The extra echo components that result from this productive process do not themselves have any meaning on their own, however, and so creating lexemes for those components would not be appropriate.

	ranna	*some value*
‘series ordinal’	1	2
‘head position’	0	1
‘head relationship’	root	echo word
UD equivalent of ‘head relationship’	root	compound:redup
‘stated as’		‘banna’

Fig. 5. Breakdown of some qualifiers on the ‘combines’ statements for the hypothetical Bengali lexeme ‘rannabanna’, where IDs for lexeme, forms, and items have been omitted for brevity.

Just as Wikidata statements and qualifiers can specify that “no value” exists as an object of their relationships, so too can they specify that the special Wikibase value “some value” exists; the implication desired here is that a component is present but that no separate Wikidata lexeme exists for it. Since, unlike the elided word case, something is still being realized in the surface form, the qualifier ‘stated as’ (`pq:P1932`) on the ‘combines’ statement can provide that form. An example of “some value” with a ‘stated as’ qualifier is shown in Figure 5.

The addition of statements with "some value" may appear to some particularly inflexible, given that individual lexeme forms can have pronunciation and grammatical information and be tied to usage examples, and that these rump 'combines' statements might become particularly unwieldy if that information were admitted there. At the same time, for languages with multiple spelling conventions (where otherwise these might be handled with separate representations using different language codes), the use of the datatype that 'stated as' expects, which does not allow attaching a language code, might lead to confusion when selecting which of a number of alternatives to use. While we cannot quite counter the latter beyond suggesting a new qualifier with a new datatype exist, for the former we might contend instead that the contribution that these echo words have outside the scope of the lexeme is especially minimal and, if one so desires, can be derived from the word it echoes without considerable difficulty.

4 Conclusions

Wikidata lexicographical data has the potential to support storing information about the syntactic structure of multi-part lexemes, and we have provided a scheme using existing Wikidata qualifiers on an existing Wikidata property for this structured information, noting as well ways in which said scheme could be improved and potential issues in pursuing those ways. Although the examples provided herein used Universal Dependencies as a basis, this by no means is limited to that particular framework; conversion between frameworks to accommodate different use cases is just as possible with Wikidata lexemes as without them. We envision the possibility of full treebanks being constructed using Wikibase and some variant of this scheme as a starting point, as well as the introduction of new structured datatypes to better handle the sorts of connections and specifications that have been handled by this scheme–all in addition to the downstream task improvements that have the potential to benefit from resources using this scheme.

References

1. Diccionario de americanismos. Asociación de Academias de la Lengua Española Penguin Random House Grupo Editorial, Barcelona (2015)
2. Darģis, R., Auziņa, I., Bojārs, U., Paikens, P., Znotiņš, A.: Annotation of the corpus of the saeima with multilingual standards. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France (may 2018)
3. Flickinger, D., Oepen, S., Ytrestøl, G.: WikiWoods: Syntacto-semantic annotation for English Wikipedia. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010), http://www.lrec-conf.org/proceedings/lrec2010/pdf/432_Paper.pdf

4. Gantar, P., Krek, S.: Slovene lexical database. Natural language processing, multilinguality pp. 72–80 (2011)
5. Grégoire, N.: Duelme: a dutch electronic lexicon of multiword expressions. Language Resources and Evaluation **44**(1), 23–39 (2010)
6. Haverinen, K., Ginter, F., Laippala, V., Viljanen, T., Salakoski, T.: Dependency annotation of wikipedia: First steps towards a finnish treebank. In: Eighth International Workshop on Treebanks and Linguistic Theories. p. 95 (2009)
7. Nguyen, K.H., Ock, C.Y.: Using wiktionary to improve lexical disambiguation in multiple languages. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. pp. 238–248. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
8. Shudo, K., Kurahone, A., Tanabe, T.: A comprehensive dictionary of multiword expressions. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 161–170 (2011)
9. Turki, H., Vrandecic, D., Hamdi, H., Adel, I.: Using wikidata as a multi-lingual multi-dialectal dictionary for arabic dialects. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). pp. 437–442 (2017). https://doi.org/10.1109/AICCSA.2017.115
10. Vondřička, P.: Design of a multiword expressions database. The Prague Bulletin of Mathematical Linguistics **112**(1), 83–101 (Apr 2019). https://doi.org/10.2478/pralin-2019-0003, http://dx.doi.org/10.2478/pralin-2019-0003
11. Vrandecic, D.: Architecture for a multilingual wikipedia. CoRR **abs/2004.04733** (2020), https://arxiv.org/abs/2004.04733