Introducing Lexical Masks: a New Representation of Lexical Entries for Better Evaluation and Exchange of Lexicons

Bruno Cartoni, Denny Vrandečić, Daniel Calvelo Aros, Saran Lertpradit
Google
{brunocartoni, vrandecic, dcalvelo, slertpradit}@google.com

Abstract

The evaluation and exchange of large lexicon databases remains a challenge in many NLP applications. Despite the existence of commonly accepted standards for the format and the features used in a lexicon, there is still a lack of precise and interoperable requirement specifications about how lexical entries of a particular language should look like, both in terms of the numbers of forms and in terms of features associated with these forms. This paper presents the notion of "lexical masks", a powerful tool used to evaluate and exchange lexicon databases in many languages.

Keywords: lexicon, interoperability, evaluation

1. Introduction

Lexicon databases are at the core of many NLP systems. Yet their maintenance, evaluation and exchange between different systems can be cumbersome. Aside from the evaluation of the quality of individual entries, the evaluation of the consistency of the structure is a problem that is hard to tackle manually. From an interoperability perspective, several attempts have been made to homogenize the morpho-syntactic features used to represent the lexicon, but few initiatives try to provide minimal requirements concerning the internal structure of a lexicon entry and its minimal specification.

In this paper, we introduce the concept of *lexical masks* which aims at representing, in a consistent way, the expected internal structure of lexical entries. Masks are defined for each language and each part-of-speech in that language. We first describe the basic principles of *lexical masks*, how these masks have been designed for several languages, and how they have been used to automatically evaluate an existing lexicon. We also introduce different levels of specifications that account for the defectiveness of some lexical entries, but also allow some flexibility in the verification process. Finally, we describe how the masks are used in Wikidata to validate existing entries and help contributors to add entries more easily.

2. State of the Art

The interoperability of language resources has always been a challenge for the NLP community. The notion of interoperability covers two dimensions (Ide and Pustejovsky, 2010): *syntactic interoperability* and *semantic interoperability*. The first one involves the data format and communication protocol while the second one focuses on the meaningful interpretation of one system's data by another.

In the domain of machine-readable lexicons, different projects have tackled the difficult issue of unifying features to describe lexical entries, allowing a meaningful *semantic interoperability*. Specifically, projects have aimed to unify the descriptors used to label specific forms in a lexical entry. Examples include the General Ontology of Linguistic Description GOLD 2010[1] (Farrar S., 2010) and the Universal Dependency Tree-bank[2] (Zeman et al., 2017) that inherently has to define a core set of tags to represent inflected forms.

On the *syntactic interoperability* side, one of the most universally accepted models, the lemon-ontolex model,[3] describes the general structure of the lexicon (what is a lexical entry, how should it be organized, how to model and organize semantic, syntactic and morphological phenomena?). This model is largely inspired by previous models such as LIR Linguistic Information Repository (Montiel-Ponsoda et al., 2011), LMF Lexical Markup Framework (Francopoulo et al., 2006), or MILE Multilingual Isle Lexical entry (Atkins et al., 2002). More recently, OLiA Ontology of Linguistic Annotations (Chiarcos, 2012) aims at unifying the annotation terminology for linguistic phenomena (including GOLD).

Thanks to these and similar projects, it is possible to define and reuse an annotation schema (and tag set) to describe lexical forms and features of the lexicon of any language. However, we are still lacking an approach that can specify how lexical entries should look like in a specific language, i.e. determining how many forms are expected and what features are necessary to describe these forms. This lack of model (together with an efficient way to represent it) makes it hard to evaluate the completeness and the coherence of lexical entries. The model also has to be operationalized in order to be automatically applicable in an evaluation process.

In order to select a serialization to represent lexical masks, we chose ShEx (Prud'hommeaux et al., 2014) (see Section 5.), a language to describe shapes for RDF graph data. ShEx can be used to generate forms and to check data. There has been precedence by (Nielsen et al., 2019) for validating Danish Wikidata lexemes. We build upon this idea and propose to make the mask a standard representation for lexical entries in every language and for every part-of-speech. To this respect, the ShEx files are not only

[1] http://linguistics-ontology.org/gold
[2] https://universaldependencies.org/
[3] https://www.w3.org/2016/05/ontolex/

for validation (Section 4.) but also for creating input forms (Section 7.).

The present paper proposes a solution for specifying lexical masks: it shows how we can, within the existing models, define the lexical entries in detail, and how we can use existing technologies to operationalize these definitions in order to check lexical entries. This solution also lends itself to be easily shared, and all masks that we create are being published to the commons, for shared ownership and maintenance.

3. Lexical Mask

3.1. Basic Principles

Lexical masks are specifications of the requirements a lexical entry should fulfill. In particular, a mask defines:

- how many forms the entry should have to be complete;
- what features are expected for each form.

Masks are specific to part-of-speech and language. One particular part-of-speech of one particular language can have more than one mask (see Section 3.3.).

	SingNumber	**PlurNumber**
MascGender	*form1*	*form2*
FemGender	*form3*	*form4*

Table 1: Lexical Masks for Italian Adjectives

For example, Table 1 shows the specification for Italian adjective entries. It specifies that four forms are expected, and each form should have one unique combination of gender and number features (i.e. there is one form for each feature bundle: MascGender / SingNumber, MascGender / PlurNumber, FemGender / SingNumber, and FemGender / PlurNumber). We do not commit to a specific tag set, but different tag sets can be used to represent the features.

Of course, lexical entries can be (and often are) much more complex, both in terms of numbers of forms, but also in how the forms are being combined from the different available dimensions, in terms of the features used to describe these forms, and the entry in general. In the following, we review how our approach tackles this complexity.

3.2. Distinguishing entry-level and form-level features

Lexical entries are not only characterized by their forms and the features associated with the forms, but also by the feature assigned at the entry-level inherent to the entire entry. For example, the mask for Russian nouns (Table 2) shows an entry-level specification that requires the combination of animacy and gender features at the entry-level, and a set of form-level features, specifying that each form must have a case and a number feature.

Examples for entry-level features include *gender* and *animacy* for nouns, *aspect* and *transitivity* for verbs, and *degree* for adjectives.

entry level	InanimateMasc OR InanimateFem OR InanimateNeut OR AnimateMasc OR AnimateFem ...

form level	**Singular Number**	**Paucal Number**	**Plural Number**
NomCase	*form1*	*form10*	*form19*
GenCase	*form2*	*form11*	*form20*
DatCase	*form3*	*form12*	*form21*
AccCase	*form4*	*form13*	*form22*
InstCase	*form5*	*form14*	*form23*
PrepCase	*form6*	*form15*	*form24*
PartCase	*form7*	*form16*	*form25*
LocCase	*form8*	*form17*	*form26*
VocCase	*form9*	*form18*	*form27*

Table 2: Russian Nouns

3.3. Accounting for more granularity: multiple masks

The configuration of lexical entries must also provide a certain level of flexibility to account for different structures of different entries. For example, we designed two masks for German nouns: the first mask, shown in Table 3 concerns nouns that have an intrinsic gender (i.e. at the entry level) and all the case and number declensions of that noun. The second mask, given in Table 4, describes the nouns that don't have an inherent gender at the entry-level but have specific inflections per gender.

entry-level	MascGender OR FemGender OR NeutGender

form-level	**SingNumber**	**PlurNumber**
NomCase	*form1*	*form2*
AccCase	*form3*	*form4*
DatCase	*form5*	*form6*
GenCase	*form7*	*form8*

Table 3: German Nouns with Gender at the Entry Level

3.4. Accounting for more granularity: canonical masks and "silver masks"

In addition to allow multiple masks for a specific part-of-speech, we also implement "silver masks" to account for phenomena such as defectiveness. Roughly speaking, defectiveness defines cases of incomplete morphological paradigms. When the defectiveness is regular enough, it should be specified in the lexical entry requirement. "Silver masks" define "smaller" entries and are a subset of the canonical mask, such as in the case of "weather verbs" in English (e.g. *"it rains"*), that only allow for a third person singular conjugation. Unlike multiple masks, "silver masks" should be used with caution, because they are only setup for a very small subset of lexical entries.

3.5. Implicit linguistic design

Designing lexical masks for a language requires us to take decisions on the structure of the lexical entries. Our model

	SingNumber		PlurNumber	
	MascGender	**FemGender**	**MascGender**	**FemGender**
NomCase	*form1*	*form2*	*form3*	*form4*
AccCase	*form5*	*form6*	*form7*	*form8*
DatCase	*form9*	*form10*	*form11*	*form12*
GenCase	*form13*	*form14*	*form15*	*form16*

Table 4: German Nouns with Gender at the Form Level

attempts to be consistent across languages, but necessarily implies theoretical choices. In particular, the distinction between "entry-level" and "form-level" features implies some decisions. For example, "degree" features for adjectives could be considered either as a form-level feature (and an adjective mask could have a duplicated set of forms for comparative and superlative forms), or as entry-level features.

Another example of a lexical design decision is for German nouns. As described above, we set up two different masks, one for 'standard nouns' with an inherent gender at the entry level, and one for nouns that have a gender inflection at the form level. Nouns denoting professions are the canonical example. Whereas standard text books of German grammar often don't mention nouns with gender given on the form-level, our experience shows that in many applications it is more useful to specify the gender at the form-level instead of the entry-level, linking masculine and feminine forms directly when they are morphologically marked. Note also that, since the nouns in Table 4 refer to descriptions for humans (professions, titles), they only support two genders (masculine and feminine), and is lacking a neuter gender.

Similarly, paucal numbers in Russian can be seen as a pure syntactic construction influenced by the choice of numerals, or can be considered as a lexical features recorded as such in the lexicon (as expressed in the mask in Table 2).

4. Using masks for lexicon validation

As mentioned above, the evaluation of the internal structure of the lexicon (in term of consistency in the number of forms and the features used to described those forms) can be quite cumbersome. The mask model presented here is used to perform a semi-automatic evaluation of the lexicon we ingest in our database.

As shown in Figure 1, each lexicon entry of a particular language (in the example an Italian adjectival entry) is ingested through the mask. During this process, we are checking that (1) this adjectival entry has indeed four forms, and (2) that each form has one of the required unique combinations of gender and number features (e.g. we cannot have two forms that are plural and feminine).

This evaluation process will mark all the entries that are passing the masks as "structurally valid". The other entries that are not passing the masks will have to be looked at more carefully by a human rater.

The interaction of masks with entries lends itself to a felicitous process of iterative refinement. It often happens that a bunch of entries are not passing our masks not because they are incomplete, but because our masks were set up only using the canonical representation of the lexical entry, but it

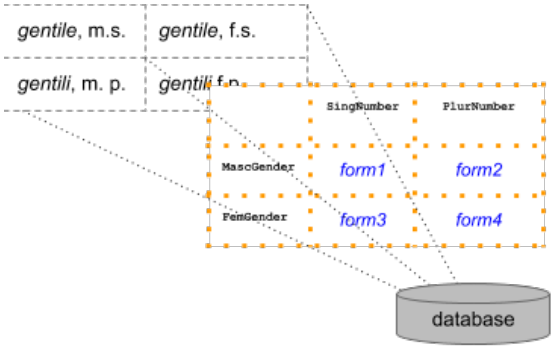

Figure 1: Evaluating a lexical entry through a mask

turned out that the entry is in fact more constrained (see Section 3.4.). For example, a number of French verbs were not passing our verbal mask. When looking at them, we realized that they were "intransitive verbs" that have only a single form for the past participle (and not the four expected forms, because they don't need to inflect). In such cases, we also create "silver masks" that are essentially subsets of the golden mask, and that "allow" the ingestion of valid entries, although not canonical. Such entries then need to be marked specifically as belonging to a more specific set, e.g. intransitive verbs, so that we can continue to select the most appropriate mask for validation.

5. Shared format: ShEx

The canonical data format of the Lemon-Ontolex ontology (McCrae et al., 2014) uses the RDF datamodel (Lassila et al., 1998). RDF is a very flexible model, but in the last few years it was increasingly recognized that in order to effectively use it, it is necessary to be able to validate the format for completeness and apply certain constraints before usage. This allows for the code using the RDF data to become much simpler. This lead to a number of proposals to describe and define constraints and completeness requirements for RDF data. The two most prominent languages currently in use for this task are SHACL (Knublauch and Kontokostas, 2017) and ShEx (Prud'hommeaux et al., 2014). In the following we will describe ShEx, as our use case uses ShEx (as described in Section 6.).

The ShEx file in Figure 3 represents the mask from Table 3.[4] In line 12, the SPARQL (Pérez et al., 2006) query is given to find all lexicographic entries the ShEx file applies to (all possible focus nodes for the shape described by

[4]This file is published at `https://www.wikidata.org/wiki/EntitySchema:E131`

the ShEX file). Below then, we see the description of the lexical entry: in line 22, we require the grammatical gender to be given at the entry level, and in lines 23ff. we see the definition of the eight individual forms that constitute a German noun as per Table 3.

The ShEx files can be used with any lexicographic data published in a data model that is compatible with the Lemon-Ontolex model. The ShEx files are all publicly available under a CC0 license (Creative Commons, 2009), and thus can be reused and modified as needed for the given use case.

The validation will ensure that all required forms are present, that the right combination of grammatical features are given throughout the forms, and that all entry-level values are set as required. Furthermore, as usual with RDF, the validation will not prevent the data from having additional annotations and markers, e.g. it will not interfere with semantic annotations on the lexical entries, or linkages between entries from different languages. The ShEx files exclusively check for the completeness of the morpho-syntactic forms of the lexical entry.

6. Use case: Validating entries in Wikidata

Wikidata (Vrandečić and Krötzsch, 2014) is a project by the Wikimedia Foundation and a sister project of Wikipedia and Wiktionary. It has the aim of creating a free knowledge base that anyone can edit. It started with the primary goal of supporting Wikipedia, and in particular it provides structured, linked data about anything that has a Wikipedia article (called a topic). In particular that means that there are already entries for many grammatical features, such as the different grammatical cases, numbers, genders, etc. In Wikidata, every topic receives a unique number, called Q Identifiers, which are language-independent. For example, the nominative case has the Q-ID `Q131105`[5], the German language `Q188`, and the paucal number `Q489410`.

Wikidata introduced the ability to add and work on lexicographic data in Wikidata in 2018. Since then, hundreds of thousands of lexical entries have been created. Wikidata publishes its data using the RDF data model. For the lexicographic data, Wikidata also follows the Lemon-Ontolex ontology, and therefore stays conceptually aligned with efforts to publicize lexicographic data on the linked open data Web (Chiarcos et al., 2013).

In 2019, Wikidata started to use ShEx to define constraints and completeness requirements for its data (Thornton et al., 2019). Although this feature was originally developed for the ontological data in Wikidata, it is also available for the lexicographic data. This allows us to enter and maintain ShEx shape files in Wikidata itself, where they can be collaboratively maintained and used to validate and check the completeness of Wikidata lexicographic data. This is the reason why we chose ShEx over SHACL to represent the constraints. In fact, we are not the first ones to follow this approach in Wikidata: (Nielsen et al., 2019) has previously

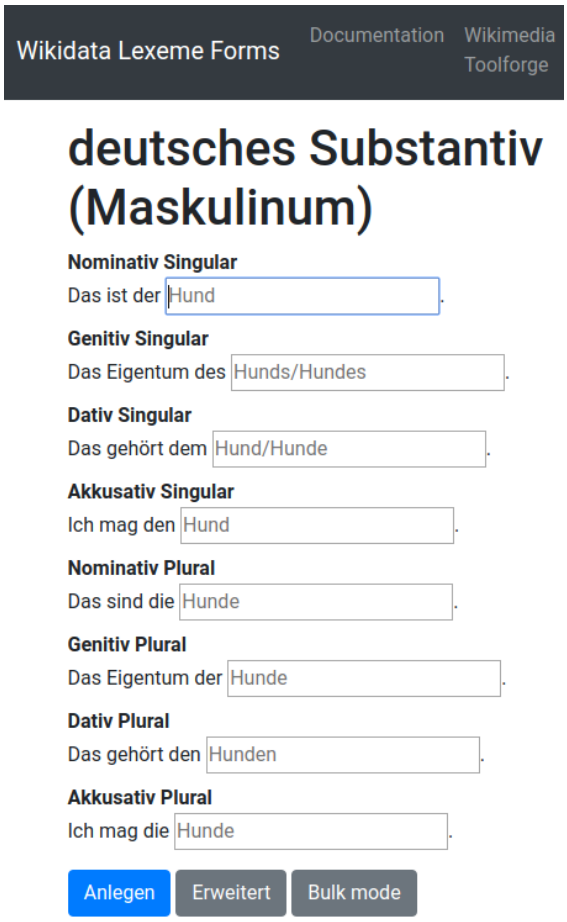

Figure 2: A screenshot of Lucas Werkmeister's tool "Wikidata Lexeme Forms" for entering forms, available at `https://tools.wmflabs.org/lexeme-forms`

created ShEx files for Danish lexicographic entries. We extend the method to generalize to more languages and use cases, with the goal to be useful well beyond the initial Wikidata use case.

Wikidata is developing its platform and infrastructure to support ShEx files in a wide range of use cases across Wikidata. Most importantly for us, we can use the files we publish to validate existing lexicographic entries. This allows for the large semi-automatic validation of the crowdsourced entries in Wikidata, and thus provides a feedback loop for the community to see the quality of their entries. They can get a generated list of all entries that do not fulfill the constraints described in the ShEx files, and then decide case by case how to handle the data (i.e. whether it is a valid exception, whether it requires an alternative or silver mask, or whether the entry needs to be improved).

7. Possible usage: mask for UI

Besides using the lexical masks for validation, we also use them for a second use case: data entry editing.

Having a mask to enter the different forms and the required entry-level data can vastly improve not only the speed of entry and the quality of the entered data, but – and that might be most important – also the satisfaction of the contributors. Since Wikidata is a crowdsourced platform, the

[5]Wikidata Q-Identifiers can be resolved both for human and machine consumption using Linked Open Data access patterns, in particular by prefixing the correct namespace. So the Website about the Q-ID `Q131105` can be found at `https://www.wikidata.org/wiki/Q131105`

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>

# Standard German noun
# German noun with gender at the entry level and
# eight inflected forms in case and number

# find all entries that are German (Q188) and nouns (Q1084)
# SELECT ?focus {?focus dct:language wd:Q188;wikibase:lexicalCategory wd:Q1084}

start = @<de-n>

<de-n> {
  dct:language [ wd:Q188 ] ; # German (Q188)
  wikibase:lexicalCategory [ wd:Q1084 ] ; # Noun (Q1084)
  wikibase:lemma [ @de ] ;
  # Grammatical Gender (P5185):
  # Male (Q1775415), Neutral (Q499327), Female (Q1775461)
  wdt:P5185 [ wd:Q1775415 wd:Q499327 wd:Q1775461 ] ;
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q131105 ]; # Nominative
      wikibase:grammaticalFeature [ wd:Q110786 ]; # Singular
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q146233 ]; # Genitive
      wikibase:grammaticalFeature [ wd:Q110786 ]; # Singular
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q145599 ]; # Dative
      wikibase:grammaticalFeature [ wd:Q110786 ]; # Singular
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q146078 ]; # Accusative
      wikibase:grammaticalFeature [ wd:Q110786 ]; # Singular
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q131105 ]; # Nominative
      wikibase:grammaticalFeature [ wd:Q146786 ]; # Plural
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q146233 ]; # Genitive
      wikibase:grammaticalFeature [ wd:Q146786 ]; # Plural
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q145599 ]; # Dative
      wikibase:grammaticalFeature [ wd:Q146786 ]; # Plural
  };
  ontolex:lexicalForm {
      wikibase:grammaticalFeature [ wd:Q146078 ]; # Accusative
      wikibase:grammaticalFeature [ wd:Q146786 ]; # Plural
  };
}
```

Figure 3: ShEx file for the mask for German nouns

satisfaction of the contributors has a direct effect on how much data they enter over time, and on how long they remain active members of the Wikidata community.

By encoding the possible forms and entry-level features beforehand, we significantly reduce the contributors' mental load while adding data. They don't have to switch between focusing on a lexeme and the actual individual forms, and the possible forms a lexeme in that language and part-of-speech might have. They don't need to create the different forms and their respective features, but merely enter the actual forms into a form-based user interface.

Note that this does not preclude exceptions. They still can be represented in the datamodel and entered through Wikidata's traditional user interface. In Wikidata, these exceptions can then be explicitly marked, e.g. as an intransitive verb or a plurale tantum, etc.

The editing forms are created automatically from the released ShEx files, using existing Wikidata infrastructure developed for form-based entry of lexicographic data (see Figure 2).

8. Conclusion and Future Plans

In this paper we presented *lexical masks* as a mechanism to share specifications of lexical entries. We showed how we can use new, but existing technologies such as ShEx to easily apply the specifications of the lexical masks to a large lexicon with small effort, in order to validate the lexical entries, and find exceptions, incomplete entries, and areas where more work is needed.

We have published the ShEx files for different parts-of-speech and languages in Wikidata. We hope that by donating these to the public domain, we will maintain and extend them together with the wider lexicographic community.[6]

We are also working on a process where we can develop more masks for languages we do not have expertise in with the wider community. We would offer the skills needed to formalize the masks as ShEx files, and work with contributors who have language expertise in creating the lexical masks. We hope that this way we can formalize a wider set of languages than has been available so far, thus expanding the reach of these technologies to many more languages.

On a more technical side, we also plan to extend the mask model with more internal checks. For example, we'd like to include simple morphological patterns inside the forms of the mask, in order to account for generic morphological paradigms. This will also allow to automatically generate regular inflected forms, speeding up even more the editing process.

9. Acknowledgement

The first implementation of the mask infrastructure was done by Adnan Öztürel. We thank Lydia Pintscher and Lucas Werkmeister of Wikimedia Deutschland for their support.

[6]The full list of published masks is available at `https://www.wikidata.org/wiki/Wikidata:Lexical_Masks` and is expected to further grow.

10. Bibliographical References

Atkins, S., Bel, N., Bertagna, F., Bouillon, P., Calzolari, N., Fellbaum, C., Grishman, R., Lenci, A., MacLeod, C., Palmer, M., Thurmair, G., Villegas, M., and Zampolli, A. (2002). From resources to applications. designing the multilingual ISLE lexical entry. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Lexical Linked Data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Chiarcos, C. (2012). Ontologies of linguistic annotation: Survey and perspectives. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 303–310, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Creative Commons. (2009). CC0. `https://creativecommons.org/share-your-work/public-domain/cc0/`.

Farrar S., L. D. (2010). An owl-dl implementation of gold. In Witt A. et al., editors, *Linguistic Modeling of Information and Markup Languages. Text, Speech and Language Technology*, volume 41. Springer.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). *Lexical markup framework (LMF)*. Wiley.

Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.

Knublauch, H. and Kontokostas, D. (2017). Shapes constraint language (shacl). *W3C Recommendation*, 20.

Lassila, O., Swick, R. R., et al. (1998). Resource Description Framework (RDF) Model and Syntax Specification.

McCrae, J., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Montiel-Ponsoda, E., De Cea, G. A., Gómez-Pérez, A., and Peters, W. (2011). Enriching ontologies with multilingual information. *Natural language engineering*, 17(3):283–309.

Nielsen, F. Å., Thornton, K., and Gayo, J. E. L. (2019). Validating Danish Wikidata lexemes. In *SEMANTICS 2019 Posters and Demos*.

Pérez, J., Arenas, M., and Gutierrez, C. (2006). Semantics and complexity of sparql. In *International semantic web conference*, pages 30–43. Springer.

Prud'hommeaux, E., Labra Gayo, J. E., and Solbrig, H. (2014). Shape expressions: an rdf validation and transformation language. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 32–40. ACM.

Thornton, K., Solbrig, H., Stupp, G. S., Gayo, J. E. L., Mietchen, D., Prud'hommeaux, E., and Waagmeester, A. (2019). Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation. In *European Semantic Web Conference*, pages 606–620. Springer.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57:78–85.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.