The Icelandic Word Web: A Language Technology Focused Redesign of a Lexicosemantic Database

Hjalti Daníelsson, Jón Hilmar Jónsson, Þórður Arnar Árnason, Alec Shaw, Einar Freyr Sigurðsson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies, Iceland
{hjalti.danielsson,jon.hilmar.jonsson,thordur.arnason,alec.shaw,
einar.freyr.sigurdsson,steinthor.steingrimsson}@arnastofnun.is

Abstract

The new Icelandic Word Web (IW) is a language technology focused redesign of a lexicosemantic database of semantically related entries. The IW's entities, relations, metadata and categorization scheme have all been implemented from scratch in two systems, OntoLex and SKOS. After certain adjustments were made to OntoLex and SKOS interoperability, it was also possible to implement specific IW features that, while potentially nonstandard, form an integral part of the Word Web's lexicosemantic functionality. Also new in this implementation are access to a larger amount of linguistic data, a greater variety of search options, the possibility of automated processing, and the ability to conduct research through SPARQL without possessing a mastery of Icelandic.

1 Introduction

We introduce the new Icelandic Word Web (IW; Icel. *Íslenskt orðanet*), a language technology focused overhaul and redesign of a lexicosemantic database of semantically related Icelandic words and phrases (Jónsson, 2017). This modernization improves access to the IW's intricate systems, makes its data more malleable, enables the use of a greater variety of metadata, and allows for a new, open-ended approach to conducting research on its various elements.

The IW is the only database of its kind for the Icelandic language. Although there does exist a number of other semantic databases, e.g. Arabic WordNet (Black et al., 2006), BalkaNet (Tufis et al., 2004), EuroWordNet (Vossen, 1998), IndoWordNet (Bhattacharyya, 2010), and The MultiWordNet Project (Pianta et al., 2002), there is a strong tendency for these to be modeled on the Princeton WordNet (Princeton University, 2010), arguably one of the best known databases of semantic word relations. While comparisons might be made between the IW and the Princeton WordNet, the IW diverges considerably in its overall structure and approach to semantic relations; its structure is more fluid and its focus more on the relations between core entries rather than the complex hierarchy around them (Rögnvaldsson, 2018). The implementation of the new IW itself represents a novel application of the two models with which the IW is encoded, and may prove useful in other projects involving the encoding of lexical databases with nonstandard structures and elements.

We begin by describing the core structure of the original IW, focusing on the aspects that remained unaltered. We then move on to the details of the overhaul. We discuss the choice of implementation models, how we applied them to the IW and what benefits we derived, and how we adapted them to certain aspects of the IW that were vital to its design but could not be represented by standard model features. We subsequently describe how the redesign has increased search scope, both in terms of the amount of accessible data and of the ways in which that data may now be searched for and inspected. Lastly, we touch on the potential future development and use of the IW, now that it is in this new form.

2 Core Structure of the Icelandic Word Web

The IW is effectively composed of two separate but interconnected systems: Entries and categories.

The former, entries, contains the words themselves and their semantic relations, and forms the bulk of the IW. Entries come in many varieties: Monolexical and polylexical, unordered and ordered (including phrasemes), sourced both from reference works and primary sources, and accompanied by varying degrees of explanatory and morphosyntactic metadata (Jónsson, 2018). As is common with these types of collections, the entries do not have definitions except in cases where glosses are necessary to differentiate word forms; rather, their meanings are considered to be implicit in the relations they have to other entries or to their respective categories. The semantic relations themselves are similarly sourced both from primary sources and older reference works, with the majority being derived from the former.

The latter system, categories, contains a semantic classification scheme, and effectively functions as an ontology for the IW's entries. Unlike the entries and their relations, which are primarily derived from source material, the categories have been created and implemented over the years by the IW's past administrators. The scheme is descriptive rather than prescriptive, and is not intended to be all-encompassing; each entry may thus belong to one, none, or multiple categories. All categories have equal priority, there are no category hierarchies, and from a semantic perspective their subjects may overlap. Although categories do exist as separate entities in the IW, there are no direct category-to-category relations. They are connected only through the relations of the entities that belong to them.

The IW's primary type of semantic relation is a specific kind of parallel construction, which we will call *Pairings* for short. This relation indicates that two given entries, X and Y, have at some point appeared in a source text with the conjunction *og* (Eng. *and*) between them. Pairings are unordered by design, with a sourced *X og Y* being considered the equivalent of *Y og X*. Most of the other relation types in the original IW build in some way on Pairings, aside from a relatively small set of synonyms and antonyms whose handcrafted relations are drawn from preexisting entries in the IW's database. Pairings combine semantics and syntax, albeit with an emphasis on the former; and this amalgamated nature, coupled with their status as a cornerstone of the IW's full span of relation types, was a major design factor in the development of the new IW.

The creation of the original IW involved the work of several people, over a period of decades rather than years, collating relational information that initially described syntax and morphology but later shifted in focus to involve semantics as well, all of which culminated in a deep and complex collection of data. While the original IW is presented through a web interface[1], there is no single, fully standardized type of entry in its underlying database. Some entries may be written or encoded differently from others, some have more metadata, and in certain cases the metadata itself may also be encoded differently between entries.

In short, a direct conversion to an established format was not an option. The only way of bringing the IW to a language technology friendly format while simultaneously maintaining its breadth of data and functionality was to design and implement it in the new format almost from scratch – a process that not only allowed for greater standardization, but also for greater inclusivity of information that up until now had either been difficult to use or entirely inaccessible.

3 General Implementation

Given that the IW's original structure is divided into two separate systems whose functionality is not the same, we approached the reimplementation from the very beginning with the idea that it would not necessarily contain only one system. We therefore expected – and later found – one of the major points of complexity not to lie in how to fit the entire structure into one paradigm, but rather how well two new schemes could interact.

We modeled the new IW using two separate systems: OntoLex (McCrae et al., 2017) and SKOS (Miles and Bechhofer). Not only was each of these well suited to represent its respective part of the IW, but their point of intersection also turned out to be both fully workable – OntoLex's functionality has been designed with SKOS's interoperability in mind, so the two models mesh well when applied to the IW – and useful to model certain important and nonstandard aspects of the IW. Moreover, the use of these systems opened up the possibility of a range of new queries into the data that had not been possible until now, both through the new systemized encoding of its metadata, which made it available to users for the first time, and through the option of user-created SPARQL queries rather than fixed web site user patterns.

OntoLex was used to encode the IW's basic entities: Lemmas, their semantic relations, and per-

[1] `https://ordanet.arnastofnun.is/`

tinent morphosyntactic information. It supports complex linguistic modeling, and was the model of choice for structuring an RDF version of the Princeton WordNet. It has already been used to recreate dictionaries in such a way that they are easily integrable with certain outside resources, an important point for the IW's two systems. Moreover, it is the only data model of its kind that can reasonably be applied to a morphologically rich language such as Icelandic (Cimiano et al., 2016). OntoLex allowed us to recreate with relative ease a myriad of the original IW's features, and, moreover, it enabled us to codify data that had been present in the original IW but had not been directly available to the user. As an example, most of the new IW's monolexical entries are now accompanied by data drawn (when available and applicable) from the Database of Icelandic Morphology (DIM) (Bjarnadóttir et al., 2019). The data cover not only each entry's lemmatized form but also its various inflectional forms and associated morphosyntactic features as well. This enables users to conduct both context-based searches for inflectional forms, and searches based on the morphosyntactic features themselves. These include gender, case, number, voice, mood, tense, person, and definiteness.

Additionally, certain entries (both mono- and polylexical) are labeled as exclamations, conjunctions, prepositions, numerals, set phrases and proper nouns, to the degree that the IW's original data allows. The encoded data for polylexical entries is somewhat more sparse than for monolexical ones, partly to save space and avoid reduplicating data. However, a valuable new feature of the IW is interlinking: Wherever possible we have added to each single word of a polylexical entry a link to that word's corresponding monolexical entry. This new feature grants access by proxy to the word's morphosyntactic data, removing the need to reduplicate all that information in the polylexical entry, and overall greatly increases both the accessibility and interrelatedness of the IW's data.

SKOS, a popular RDF-based ontology model, was used to encode the IW's categories. While OntoLex has a great variety of various encoding options but a stringently ordered design dictating their use, the base version of SKOS has a comparatively smaller range of options but is far more malleable, a fact that makes it well-suited for the IW. Instead of being a standalone entity with a complete and systematic internal structure, the IW's category system draws heavily on the content of its other system of entities – categories are only created and put into use if existing lemmas support them – and there is a great deal of commingling and cross-referencing between the categories and lemmas, which means that the category system needs to be represented by a model that does not require all its entities to be discrete. Using SKOS, modeling the categories was a straightforward process. Where SKOS really comes into play is at the point where the IW's two systems – and hence these two models – intersect.

4 Model Convergence

While OntoLex is an ideal option for representing the IW's complex grammar, it is unable to encapsulate certain other aspects of the IW's design. Most notably, OntoLex cannot comfortably represent semantic relationships such as the Pairings noted in the previous section, nor can it encode entities at all if, as is sometimes the case in the IW, they do not have an ontological connection to at least one category. SKOS, meanwhile, may be used to implement both these features but cannot store the entities themselves, which must be kept in OntoLex if we are to hold on to that linguistic modeling mentioned earlier.

Our solution was to develop a separate conceptual layer that hovers between these two models and serves as an intermediary. The change can be seen in Figures 1 and 2 below.

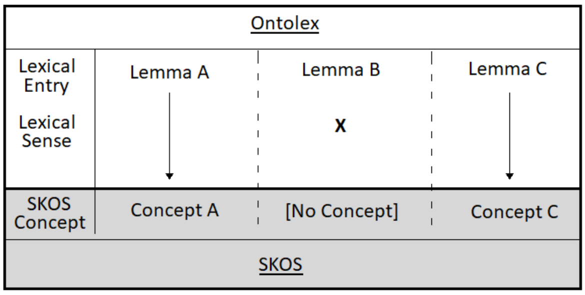


Figure 1: Potential IW implementation, without adjustment.

Figure 1 shows what the IW would be like if implemented directly in OntoLex and SKOS, without taking into account the aforementioned issues. The IW's entities would be encoded in OntoLex as Lexical Entries, while its categories would be encoded in SKOS as Concepts. The two would then be connected by an OntoLex relation called

Lexical Sense. Entities that did not belong to an IW category would not have this kind of relation, which would render them invalid in OntoLex. Moreover, the IW's Pairings relation would need to be directly between Lexical Senses, and although OntoLex does support a number of sense-to-sense relations, none of them are a suitable fit for our purpose.

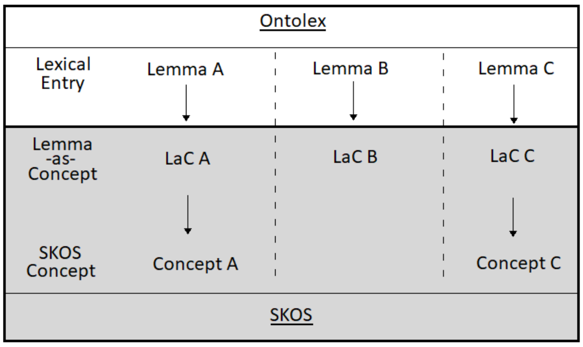


Figure 2: The new IW's actual implementation in OntoLex and SKOS.

Figure 2, on the other hand, shows our final implementation of the IW in OntoLex and SKOS, where these issues are taken into account. Here, we have added the separate conceptual layer. Note the replication of entities: After we have encoded every one of them in OntoLex as Lexical Entries, we *mirror* them in SKOS as Concepts. We call these mirrored entities *Lemma-as-Concept*, or LaC for short (*Lemma* being a more direct translation of the IW's Icelandic term for entities, *Fletta*) (Jónsson, 2017). From the viewpoint of OntoLex, LaCs serve as connectors to an ontology. From a SKOS viewpoint, LaCs are effectively a new category layer where each respective unit represents exactly one entity: The Lemma in question. In those cases where there does exist an actual category, the LaC merely functions as its subset.

Not only does this ensure that we always have the Lexical Entry/Concept connection mandatory for sustaining each Lexical Sense, but it also allows us to encode Lemma Pairings by using the SKOS *senseOf* keyword to relate LaCs as appropriate. In implementing this separate layer and its functionality, we have thus avoided creating nonstandard keywords that might have otherwise complicated the IW's use, and have confined any somewhat atypical use of the models to a clearly delineated section of our system, while simultaneously maintaining vital core functionality of the original IW. If new types of semantic word relations were to be added to the IW, they could comfortably be fitted into this layer.

5 Data Accessibility and Augmentation

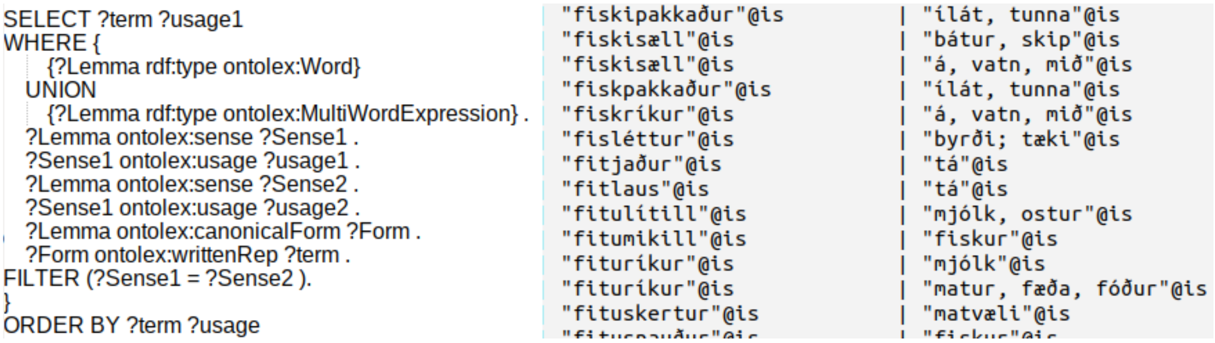
In terms of existing entity relations, the new IW has greatly increased their scope. The original IW, which is accessible through a bespoke web site, contains one fundamental semantic relation – Pairings – and three ancillary ones (Synonyms, Near-Synonyms and Antonyms), plus a half-dozen derived relations that build on these. While these relations could often be highly informative, both in content and presentation, the precise nature of each relation was fixed and could not be altered by the user. Search functionality, likewise, was simple and clean but unmalleable, with a focus on textual searches for entities. Parameters both for the search and the relations themselves could not be altered, and the results could not be exported for further examination. The original data was stored in multiple tables in a database backend behind the web site, and was either not accessible except through the web site's search options or, in the case of morphosyntactic metadata, not accessible to regular users in any way.

The new IW, by contrast, effectively offers a limitless variety of potential searches. It is stored in a single RDF file accessible directly through CLARIN[2] under a CC BY 4.0 license. We have encoded only the Pairings and ancillary relations into the system, and the derived relations are not formally encoded in the system.

Instead, everything may now be produced through queries written in SPARQL, and the raw data itself may be viewed at will as needed. The sample query in Figure 3 shows a search for all words and phrases whose written form, irrespective of grammatical categorization, has more than one "gloss", or explicitly mentioned definition. The SPARQL code is on the left, and the results on the right, with the results' left-hand column listing Word Web entries and the right-hand one listing their corresponding glosses.

These queries extend to practically every attribute encoded into the IW, including all those listed in the chapter on general implementation. So long as the user can formulate their intent into a valid SPARQL query, it may be applied to the IW's data. (This includes LaCs, which may be trivially folded into SPARQL queries.)

[2] `http://hdl.handle.net/20.500.12537/69`



```
SELECT ?term ?usage1
WHERE {
    {?Lemma rdf:type ontolex:Word}
  UNION
    {?Lemma rdf:type ontolex:MultiWordExpression} .
  ?Lemma ontolex:sense ?Sense1 .
  ?Sense1 ontolex:usage ?usage1 .
  ?Lemma ontolex:sense ?Sense2 .
  ?Sense1 ontolex:usage ?usage2 .
  ?Lemma ontolex:canonicalForm ?Form .
  ?Form ontolex:writtenRep ?term .
FILTER (?Sense1 = ?Sense2 ).
}
ORDER BY ?term ?usage
```

```
| "fiskipakkaður"@is        | "ílát, tunna"@is
| "fiskisæll"@is            | "bátur, skip"@is
| "fiskisæll"@is            | "á, vatn, mið"@is
| "fiskpakkaður"@is         | "ílát, tunna"@is
| "fiskríkur"@is            | "á, vatn, mið"@is
| "fisléttur"@is            | "byrði; tæki"@is
| "fitjaður"@is             | "tá"@is
| "fitlaus"@is              | "tá"@is
| "fitulítill"@is           | "mjólk, ostur"@is
| "fitumikill"@is           | "fiskur"@is
| "fituríkur"@is            | "mjólk"@is
| "fituríkur"@is            | "matur, fæða, fóður"@is
| "fituskertur"@is          | "matvæli"@is
| "fitusnauður"@is          | "fiskur"@is
```

Figure 3: Word Web SPARQL query and corresponding output.

6 Conclusion and Future Developments

By its encoding in a publicly accessible form, the new IW reduces the barriers to entry for anyone wishing to make use of its stores of information. It also encodes that information in such a way that far more of it is accessible and usable for research, while maintaining, wherever possible, an adherence to official standards that ensure the IW's functionality is well-documented and comparable to that of any other models encoded using those same standards. On those occasions where that adherence is not possible or practical, we have tried to ensure that non-standard use is properly documented, kept to a minimum, and contained within a specific, clearly-defined part of the IW.

The IW's data storage is kept current, with new information added on a regular basis. As noted earlier, updates of existing data will grant the IW even greater usability. In addition, the models in which the IW is encoded support a range of potential information such as bilingualism and phonetics that, although not currently a part of the IW, may now be added to a system designed to store and handle this kind of data.

Overall, the IW is a deep and extensive system that models varying degrees of semantic relations between single- and multiword lemmas, drawing its information not from third party schemas and design, but rather directly from first-party sources. There is ample reason to think that the IW will be relevant to any number of research projects in the future, particularly now that it has been redesigned and reimplemented with depth of reach and ease of access in mind.

Acknowledgements

This work is supported by the Language Technology Programme for Icelandic 2019–2023, funded by the Icelandic government.

References

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3785–3792, Valletta, Malta.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the Arabic WordNet project. In *Proceedings of the third International WordNet Conference (GWC-06)*, pages 295–299, Seogwipo, Korea.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon model for Ontologies: Community Report. *W3C Ontology-Lexicon Community Group*.

Jón Hilmar Jónsson. 2017. Frá orðabók að orðaneti. In Ásta Svavarsdóttir, editor, *Bundið í orð*, pages 1–26. The Árni Magnússon Institute for Icelandic Studies, Reykjavík.

Jón Hilmar Jónsson. 2018. Íslenskt orðanet: Tekstbasert kartlegging og presentasjon av leksikalske relasjoner. *Nordiske Studier i Leksikografi*, 14:1–17.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex 2017 conference*, pages 587–597, Leiden, the Netherlands.

Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System Reference. *W3C Recommendation, year=2009, publisher=World Wide Web Consortium*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Princeton University. 2010. `https://wordnet.princeton.edu/` About WordNet.

Eiríkur Rögnvaldsson. 2018. Íslenskt orðanet: a treasure for writers and word lovers. *LexicoNordica*, 25:313–328.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology*, 7(1–2):9–43.

Piek Vossen. 1998. Introduction to EuroWordNet. In *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, pages 1–17. Springer.