UNIVERSITY OF COPENHAGEN
DEPARTMENT OF NORDIC STUDIES & LINGUISTICS
CENTRE FOR LANGUAGE TECHNOLOGY

DET DANSKE
SPROG- OG
LITTERATURSELSKAB

The new DanNet

DanNet, the Danish WordNet, is a semantic language resource which links the many word senses of the Danish language in various ways.

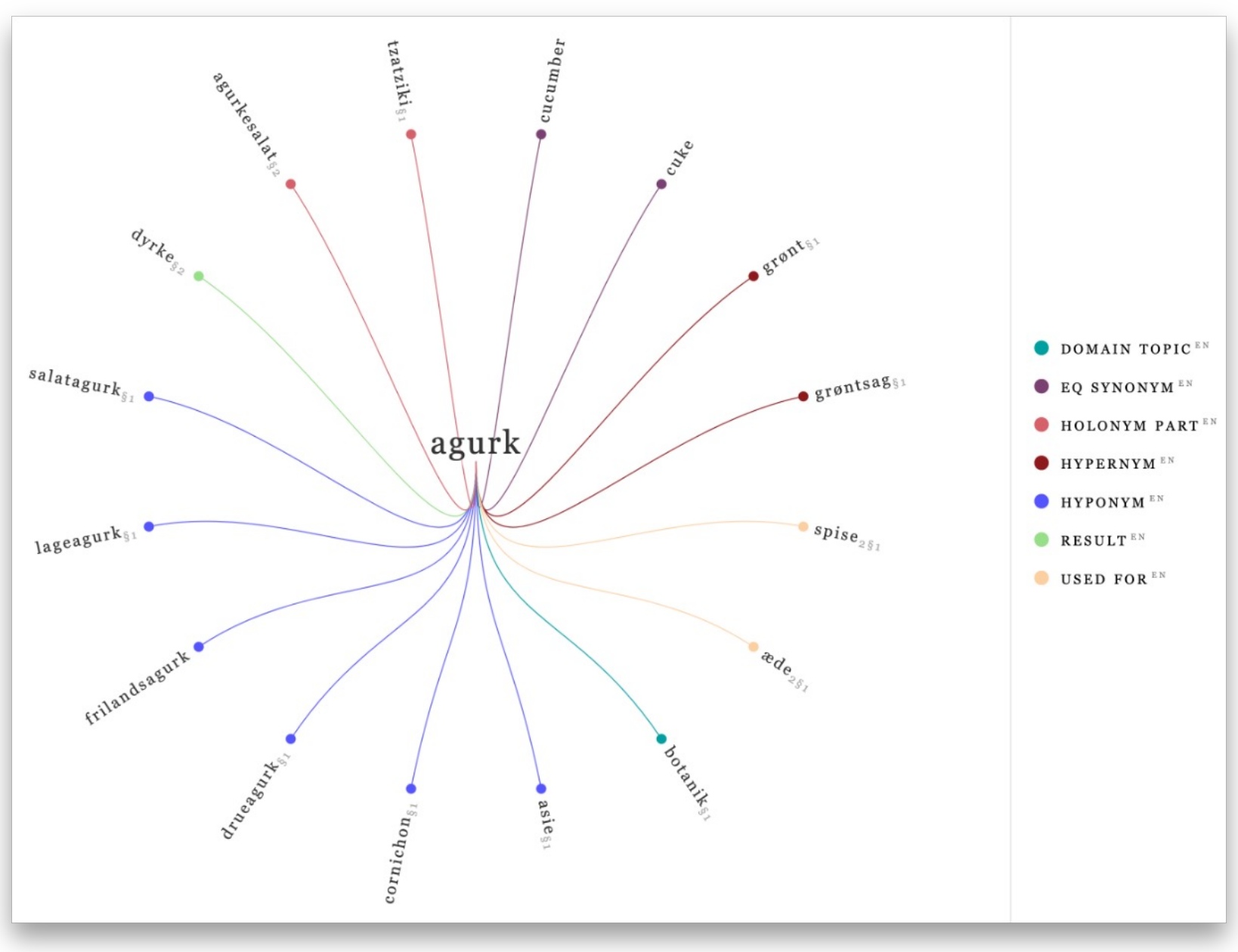
In 2023, an entirely new version of DanNet was released, containing major changes to the language resource along with a new web presence at wordnet.dk/dannet*. The project is a collaborative effort of Det Danske Sprog- og Litteraturselskab and the Centre for Language Technology.*

About WordNets

A WordNet is a lexico-semantic network graph which shows how the **concepts** of a language relate to other concepts through semantic relations. One can also think of a WordNet as a kind of machine-readable dictionary.

Unlike a typical dictionary, definitions aren't central. Instead, the **relations** between words (grouped by sets of synonyms — or **synsets**) are key. For example, you can see that a Swiss willow is a kind of bush, that a gazebo is located in a garden, that "fiberdrys" is for eating, and that cakes are typically produced by baking and usually made from flour and sugar.

DanNet currently contains **~70,000** concepts (synsets) connected by **~500,000** semantic relations.



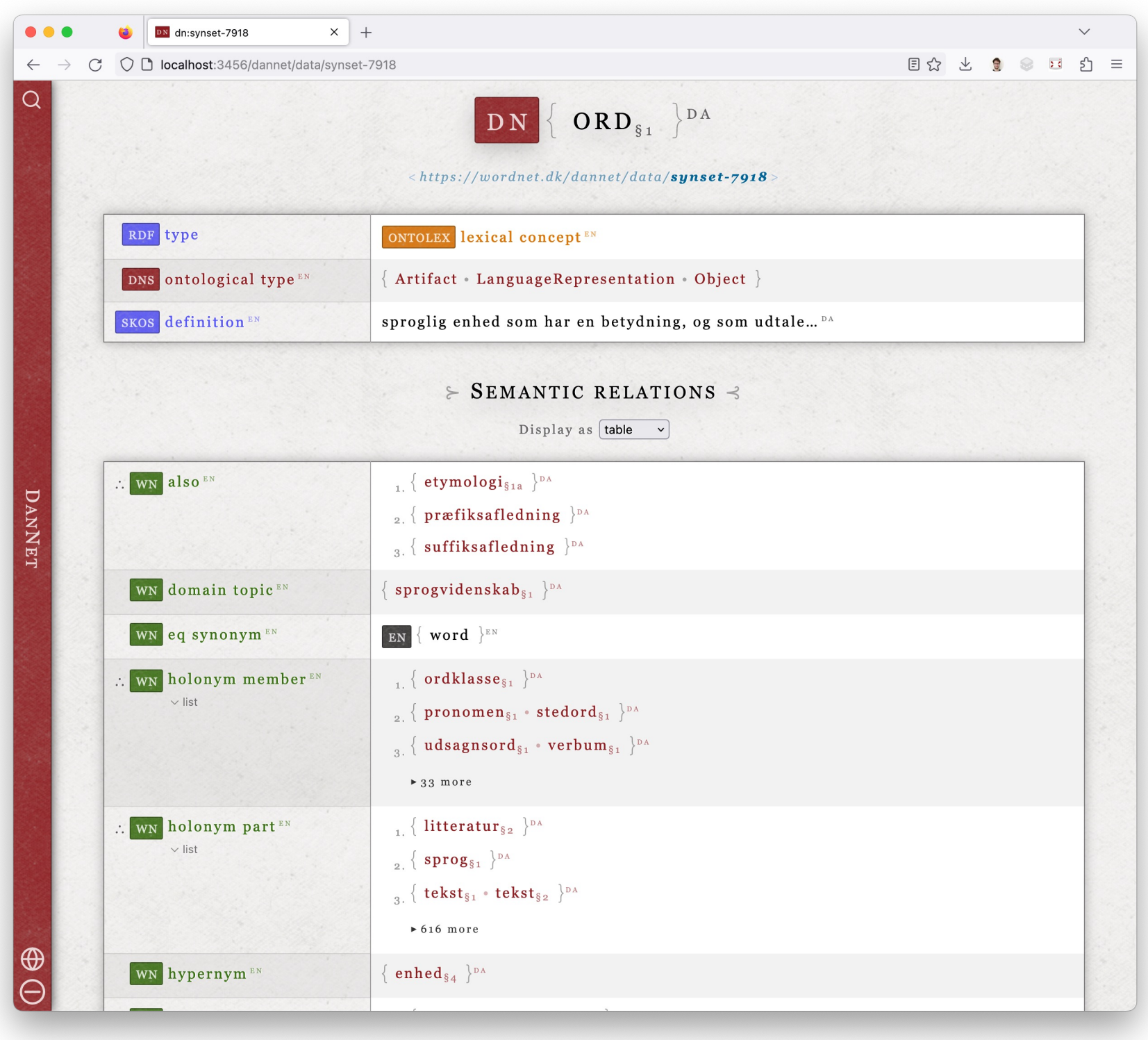
wordnet.dk/dannet/data/synset-543
The new DanNet website can also be used to show dynamic illustrations of the relations present in DanNet.

Standards-based

As of 2023, DanNet is now completely standards-based and RDF-native:

- Our data uses the **Ontolex** standard with WordNet relations from the **Global WordNet Association**.
- In addition to the work done on DanNet itself, we have also contributed to the Global WordNet Association's **RDF schema**.
- There is a new **bilingual** home available at the URL wordnet.dk/dannet, where the DanNet RDF graph can be browsed along with several linked datasets.
- The DanNet IDs all dereference as actual **RDF resources** which may be accessed in a browser in accordance with the best practices for **linked data**.

In practice: you can now inspect any DanNet resource by simply copy-pasting its full ID — i.e. URL — into the address bar of a web browser to get an interactive view.



DN { ORD$_{§1}$ }^{DA}

<https://wordnet.dk/dannet/data/synset-7918>

RDF type	ONTOLEX lexical concept^{EN}
DNS ontological type^{EN}	{ Artifact • LanguageRepresentation • Object }
SKOS definition^{EN}	sproglig enhed som har en betydning, og som udtale…^{DA}

SEMANTIC RELATIONS

Display as table

∴ WN also^{EN}	1. { etymologi$_{§1a}$ }^{DA} 2. { præfiksafledning }^{DA} 3. { suffiksafledning }^{DA}
WN domain topic^{EN}	{ sprogvidenskab$_{§1}$ }^{DA}
WN eq synonym^{EN}	EN { word }^{EN}
∴ WN holonym member^{EN} list	1. { ordklasse$_{§1}$ }^{DA} 2. { pronomen$_{§1}$ • stedord$_{§1}$ }^{DA} 3. { udsagnsord$_{§1}$ • verbum$_{§1}$ }^{DA} ▸ 33 more
∴ WN holonym part^{EN} list	1. { litteratur$_{§2}$ }^{DA} 2. { sprog$_{§1}$ }^{DA} 3. { tekst$_{§1}$ • tekst$_{§2}$ }^{DA} ▸ 616 more
WN hypernym^{EN}	{ enhed$_{§4}$ }^{DA}

wordnet.dk/dannet/data/synset-7918
Every resource in the new DanNet dataset has its own web presence. This also includes metadata and schemas.

New additions

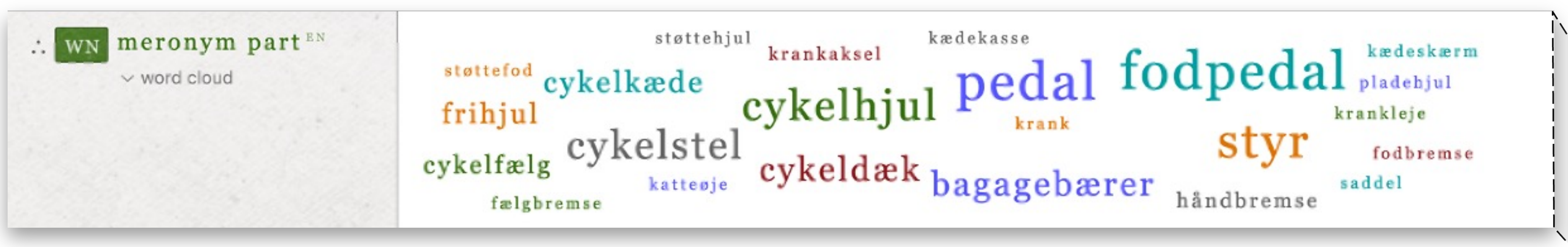
The DanNet dataset has been expanded and many dataset inconsistencies or other undesirable properties have been cleaned up. Furthermore,

- An additional **~5000** new concepts have been added, mostly adjectives.
- An additional **~5000** new links between DanNet and the Open English WordNet (OEWN), or via the Collaborative Interlingual Index (CILI).
- Most words and senses have been linked to "Den Danske Ordbog" (DDO) via our new **dns:source** relation. Many word frequencies from DDO are also included.

RDF vs. tables

DanNet and other WordNets have long been published as RDF. However, in DanNet's case, the **RDF/XML** serialisation had primarily been used as a method of distribution to third parties, while internally the DanNet graph had been modelled as tables within a traditional **SQL** database. These two models were only connected in a limited fashion, requiring adapter code.

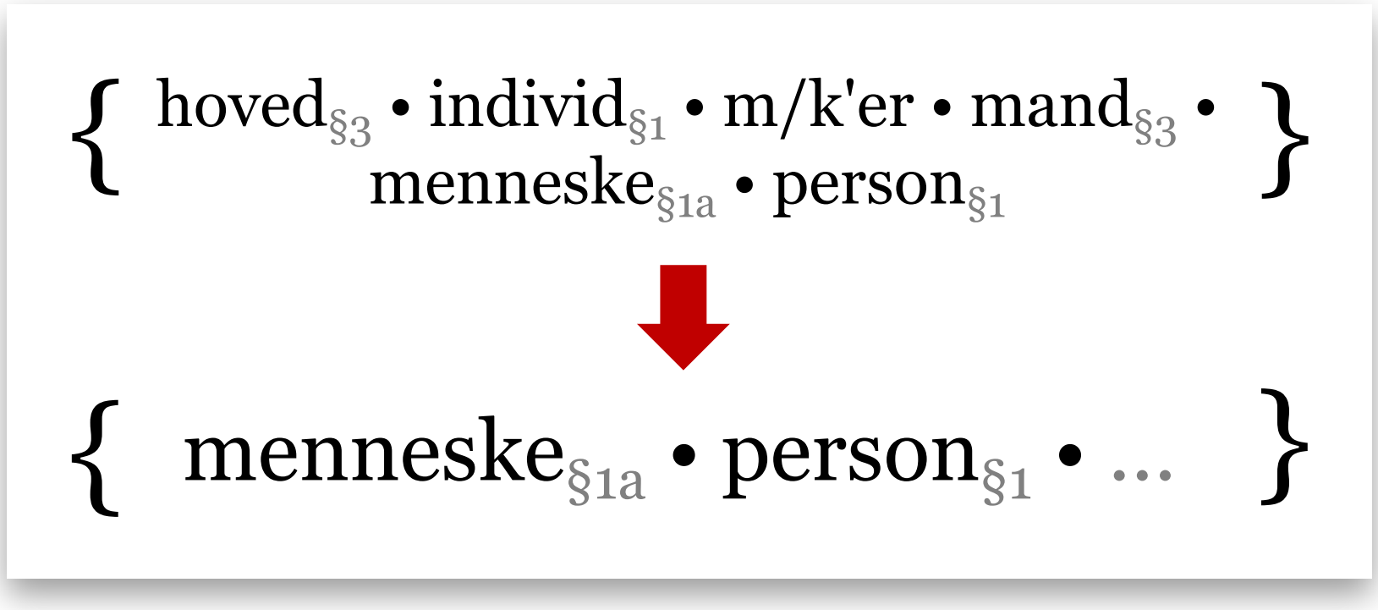
While using a more widespread data model such as the relational model has clear benefits, combining it with graphs is not as ergonomic due to the differing perspectives of the two data models. **The semantic web** stack (RDF, SPARQL, ...) is a better fit. It is quite mature after several decades of standards development.



Dynamic inline visualisations
Individual one-to-many relations can be visualised in place, e.g. as an interactive word cloud (above: bicycle parts).

The labels used in DanNet have also been greatly revised. The part of the label denoting the section of DDO where a sense may originally have been pulled from is now clearly marked using the same format as DDO, e.g. the sense label of the synset **{ øvrighed$_{§1a}$ }** is also to be found in section *1.a* of the corresponding entry on the DDO website.

The longest synset labels now also come with abbreviated **dns:shortLabel** versions which contain *only* the most canonical words of the synset.



{ hoved$_{§3}$ • individ$_{§1}$ • m/k'er • mand$_{§3}$ • menneske$_{§1a}$ • person$_{§1}$ }

{ menneske$_{§1a}$ • person$_{§1}$ • … }

wordnet.dk/dannet/data/synset-2250
The longer synset labels now also come with abbreviated variants containing *only* the most canonical words.

Dataset changes

Aside from the core DanNet dataset, several companion RDF datasets are available for download, e.g.

- Det Centrale Ordregister (COR)
- Det Danske Sentimentleksikon (DDS)
- DanNet-style labels for the Open English WordNet (OEWN)

Additionally, the DanNet CSV download is now **CSVW** and includes metadata describing the contents of the columns. However, the canonical version of the DanNet dataset will henceforth be the new RDF version, which includes the entire dataset. The RDF dataset is also better suited for immediate querying using **SPARQL**.

The DanNet dataset itself has been relicensed as **CC BY-SA 4.0**, while the source code of the DanNet project is available under the **MIT licence**.

As of 2023, both the core dataset as well as the companion datasets have frequent updates:

github.com/kuhumcst/DanNet/releases

Every past release along with the source code used to generate the data can be found at the above URL.

Serialised graphs

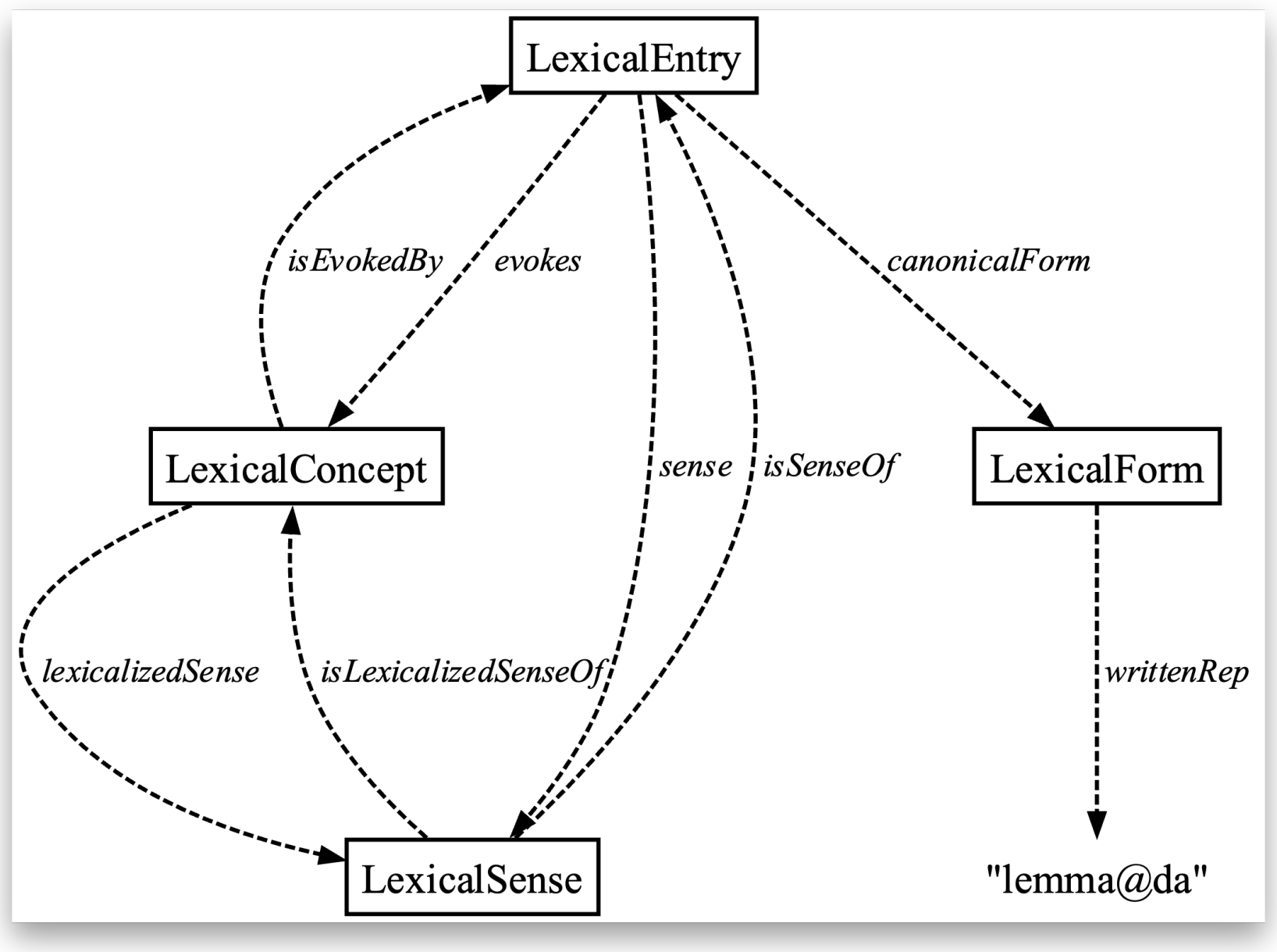
The decision to go RDF-native was based on some of these thoughts:

1. WordNets are simply network graphs.
2. Graphs can be decomposed into triplets.
3. RDF is a graph abstraction using triplets as its fundamental data structure.

The optimal data representation for DanNet allows it to be compatible with the RDF standard, decomposable into triplets, and represented as a graph in any context, i.e. at the **application level** *or* at the **database level**.

In addition, the data can always be serialised and loaded into a compatible application or be integrated with other RDF knowledge graphs, e.g. the **Princeton WordNet** or its successor: **the Open English WordNet**, the latter also using the Ontolex standard.

DanNet now defaults to the **Turtle** (.ttl) serialisation format, rather than the more archaic RDF/XML.



w3.org/2016/05/ontolex
The Ontolex standard provides a structure for the new DanNet. In Ontolex, a synset is known as a *LexicalConcept*.

```
SELECT DISTINCT ?slang
WHERE {
  ?sense lexinfo:register lexinfo:slangRegister .
  ?word  ontolex:sense ?sense ;
         ontolex:canonicalForm ?form .
  ?form  ontolex:writtenRep ?slang .
}
```

Querying DanNet
The SPARQL query above demonstrates how to retrieve slang word lemmas from DanNet using a graph pattern.

A bit of history

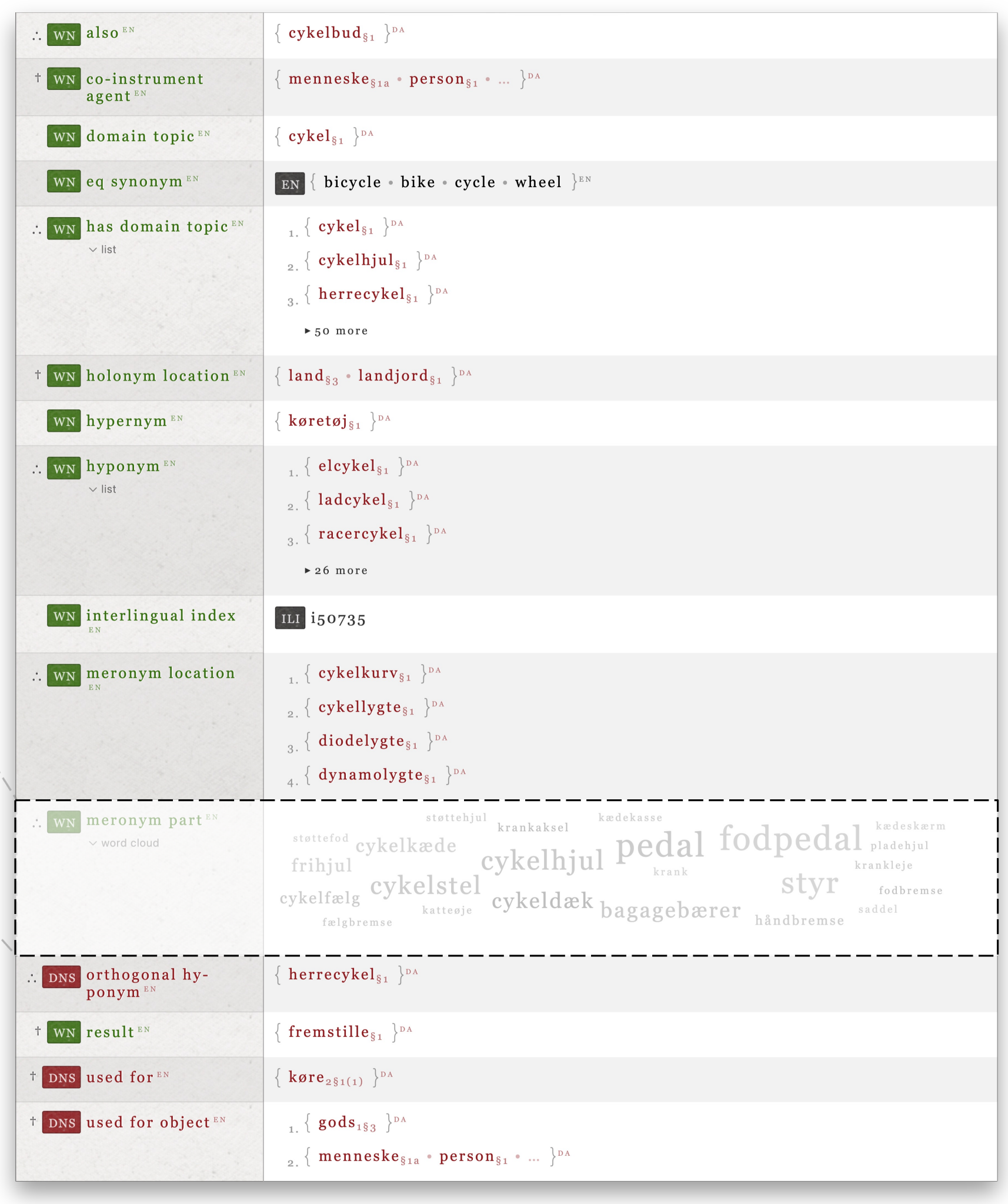
DanNet was originally released in **2009**. It has been maintained and extended until today with the help of funding from many different sources.

The **original paper** introducing DanNet was written by Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen, and Henrik Lorentzen in 2009, published in the journal *Language Resources and Evaluation.*

This paper, entitled *"DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary"*, details how DanNet originally came about as a resource derived from **Den Danske Ordbog** inspired by concepts taken from the **EuroWordNet** project. More information may be found at:

cst.ku.dk/projekter/dannet

This latest release has been funded by *Carlsberg Fondets infrastrukturpulje*. Until now, the technical side of DanNet has been maintained by DSL, but as of the 2023 release the stewardship now lies with the Centre for Language Technology (NorS, UCPH).



∴ WN also^{EN}	{ cykelbud$_{§1}$ }^{DA}
† WN co-instrument agent^{EN}	{ menneske$_{§1a}$ • person$_{§1}$ • ... }^{DA}
WN domain topic^{EN}	{ cykel$_{§1}$ }^{DA}
WN eq synonym^{EN}	EN { bicycle • bike • cycle • wheel }^{EN}
∴ WN has domain topic^{EN} list	1. { cykel$_{§1}$ }^{DA} 2. { cykelhjul$_{§1}$ }^{DA} 3. { herrecykel$_{§1}$ }^{DA} ▸ 50 more
† WN holonym location^{EN}	{ land$_{§3}$ • landjord$_{§1}$ }^{DA}
WN hypernym^{EN}	{ køretøj$_{§1}$ }^{DA}
∴ WN hyponym^{EN} list	1. { elcykel$_{§1}$ }^{DA} 2. { ladcykel$_{§1}$ }^{DA} 3. { racercykel$_{§1}$ }^{DA} ▸ 26 more
WN interlingual index^{EN}	ILI i50735
∴ WN meronym location^{EN}	1. { cykelkurv$_{§1}$ }^{DA} 2. { cykellygte$_{§1}$ }^{DA} 3. { diodelygte$_{§1}$ }^{DA} 4. { dynamolygte$_{§1}$ }^{DA}
∴ WN meronym part^{EN} word cloud	støttehjul krankaksel kædekasse støttefod cykelkæde frihjul cykelhjul pedal fodpedal kædeskærm pladehjul krankleje krank cykelfælg cykelstel cykeldæk styr fodbremse katteøje bagagebærer håndbremse saddel fælgbremse
∴ DNS orthogonal hyponym^{EN}	{ herrecykel$_{§1}$ }^{DA}
† WN result^{EN}	{ fremstille$_{§1}$ }^{DA}
† DNS used for^{EN}	{ køre$_{2§1(1)}$ }^{DA}
† DNS used for object^{EN}	1. { gods$_{1§3}$ }^{DA} 2. { menneske$_{§1a}$ • person$_{§1}$ • ... }^{DA}

wordnet.dk/dannet/data/synset-1522
Relations emanating from the synset **{ cykel$_{§1}$ }**.
The attributes in red are entirely DanNet-specific.

Future work

We are currently not planning any major additions to the core dataset, but DanNet is still being developed:

- Much of the recent work has gone into facilitating visualisations in the web interface.
- The **DanNet schema** sees occasional changes which may be reflected in the core dataset.
- Errors in the data are corrected on a regular basis.

The next challenge will be developing **guides** and **code examples** for prospective users of the dataset, e.g. how to load an RDF dataset into a graph and how to query it with SPARQL.

Discover more

If you would like to further explore the new DanNet, you may visit the following websites:

- wordnet.dk/dannet
  INTERACTIVE VERSION
- github.com/kuhumcst/DanNet
  SOURCE CODE
- cst.ku.dk/projekter/dannet
  PROJECT PAGE

Other relevant keywords or search terms:

WordNet, Ontolex, RDF, SPARQL, Turtle/.TTL, linked data

November 2023
Simon Gray
Centre For Language Technology