A *lemon* Model for the ANW Dictionary

Carole Tiberius[1], Thierry Declerck[2]

[1] Instituut voor de Nederlandse Taal, Matthias de Vrieshof 2,
2311 BZ Leiden, the Netherlands
[2] DFKI GmbH – Multilingual Technologies, 3 Stuhlsatzenhausweg,
D-66123 Saarbrücken, Germany
E-mail: carole.tiberius@ivdnt.org, declerck@dfki.de

Abstract

In this paper, we explore how we can reuse data from the ANW – an online corpus-based, scholarly dictionary of contemporary standard –, improve and optimise it by porting some of its elements into modules of the *lexicon model for ontologies (lemon)*. For the current study, the focus was set on the application of the ontolex and decomp modules, together with the associated LexInfo vocabulary in order to model the semantic and morphosyntactic features of nominal entries in the ANW.

We observe that encoding the ANW information in *lemon* has a number of advantages, including a better modularisation of the data, linking to other (lexical) data and data access using the standardised SPARQL query language.

Keywords: *lemon* model; lexical entry; semagram

1. Introduction

The Algemeen Nederlands Woordenboek (ANW) is a comprehensive online scholarly dictionary of contemporary standard Dutch, which is being compiled at the Dutch Language Institute.[1] It was set up as on online dictionary from the start and, as such, it truly represents a new generation of electronic dictionaries in the sector of academic and scientific lexicography. The dictionary focuses on written Dutch and covers the period from 1970 onwards. For a general introduction to the ANW and its features, the reader is referred to Schoonheim and Tempelaars (2010).

In this paper, we explore how we can reuse ANW data, and improve and optimise its internal formal representation by porting some of its elements into modules of the LExicon Model for ONtologies (*lemon*), using the version published as the result of the W3C Ontology-Lexica Community Group.[2] The original aim of *lemon* was to provide rich linguistic grounding for ontologies. This grounding includes the formal representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e., the meaning of these lexical entries with respect to

[1] See http://ivdnt.org/the-dutch-language-institute [accessed 18.05.2017]

[2] See https://www.w3.org/2016/05/ontolex/ [accessed 18.05.2017]

a descriptive vocabulary or an ontology (McCrae, 2012).

The main modules of *lemon* are:

- Ontology-lexicon Interface (ontolex)
- Syntax and Semantics (synsem)
- Decomposition (decomp)
- Variation and Translation (vartrans)
- Linguistic Metadata (lime)

For the current study the focus was set on the application of the ontolex and decomp modules, and we used the associated LexInfo vocabulary.[3] Ontolex is, in fact, the core module of *lemon*, describing in detail the interface between elements of a lexical entry and the conceptual or world knowledge encoded in lexicon external knowledge bases. Decomp is the module depicting how to encode elements that are a part of a multi-word or compound lexical entry. Both modules are graphically displayed in Figure 1 and Figure 3. LexInfo, building in part on the ISOcat vocabulary[4], is an ontology that was defined to provide data categories (e.g., to denote gender, number, part of speech, etc.) for the *lemon* model.

Our starting point for the study is given by a small set of representative examples of ANW lexical entries encoded in an internal XML format. Our work consisted of proposing a mapping of this XML format onto the *lemon* vocabulary, which makes use of OWL, RDF(s) and RDF constructs.[5] The objective is to investigate if the ANW data can be encoded in an improved modular manner, supporting a higher level of re-usability within the ANW dictionary environment and an improved interoperability with other data sources, especially in the context of the Linked Open Data framework.[6] At the same time, the ANW data offer an excellent source for testing the validity of the *lemon* approach for comprehensive lexicographic resources (similar to the work by El Maarouf et al. (2014), Bosque-Gil et al. (2016), Kahn et al. (2017) or Stolk (2017)) and for suggesting potential extensions.

2. Data Modelling

We started our study with the description of nominal entries in the ANW dataset, considering in the first instance a description of the semantic and morphosyntactic

[3] See http://lov.okfn.org/dataset/lov/vocabs/lexinfo for more details.

[4] See http://www.isocat.org/ for more details.

[5] See respectively https://www.w3.org/OWL/, https://www.w3.org/TR/rdf-schema/ and https://www.w3.org/RDF/

[6] See http://linkeddata.org/ and also the Linguistic Linked Open Data cloud (http://linguistic-lod.org/llod-cloud).

features of these entries. As mentioned in the introduction, the ANW is a scholarly dictionary, providing a detailed description of each lexical entry. In the dictionary, special attention is paid to words in context (combinations, collocations, idioms, proverbs), relations with other words (lexical relations like synonymy, antonymy, hypernymy, hyponymy), semantic relations within the entry (metaphor, metonymy, generalisation, specialisation) and morphological patterns, the word structure of derivations and compounds. This means that the ANW has a rich microstructure.

To model the ANW microstructure with *lemon*, we start with its core module, ontolex (ontology-lexicon interface), as depicted in black in Figure 1[7]. In red we mark the additional elements, either taken from the LexInfo vocabulary or our suggestions, for extending LexInfo in order to account for ANW features.

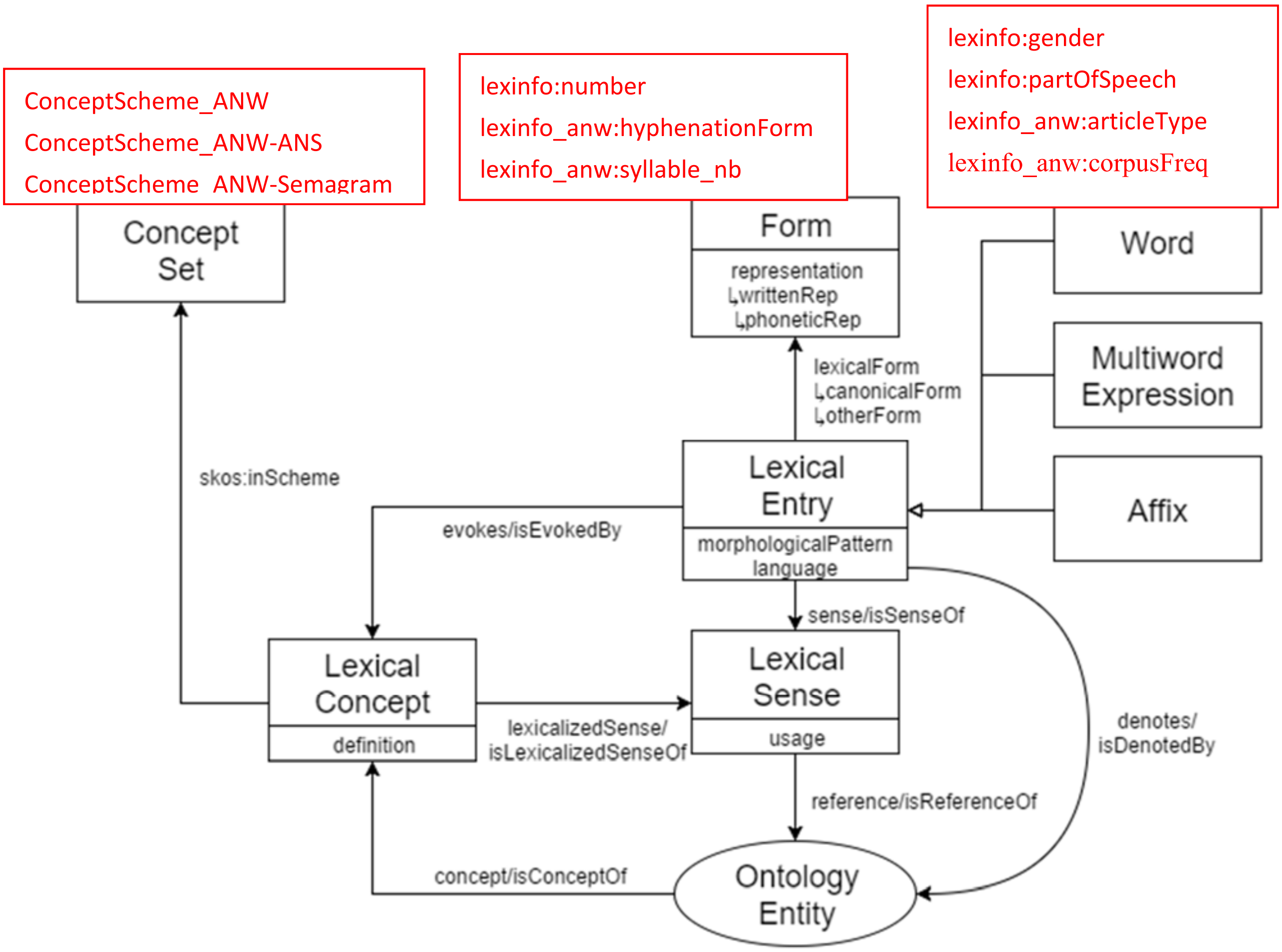


Figure 1: The ANW data model based on ontolex, the core module of *lemon*

In the remainder of this section, we discuss the *lemon* model for the ANW on the basis of an example entry, i.e. *wijn* (*'wine'*)[8]. First we discuss the morphosyntactic encoding, then the semantic encoding and we conclude with the modelling of compounds.

[7] https://www.w3.org/2016/05/ontolex/. Figure created by John P. McCrae for the W3C Ontolex Community Group.

[8] http://anw.inl.nl/article/wijn.

2.1 Encoding of morphosyntactic information

Table 1 lists the morphosyntactic features for the entry *wijn* ('wine') in the ANW. The corresponding *lemon* encoding is given in the last column. Our suggested extensions to the LexInfo vocabulary include the sub-string `anw` and are marked in red.

ANW Features	**ANW Data**	***lemon* encoding**
Lemma		
	Lemma form: wijn	:form_wijn ... ontolex:writtenRep
	Lemma type: woord	rdf:type ontolex:Word
Syntactic Category		
	Type: noun	lexinfo:partOfSpeech lexinfo:noun
	Name type: soortnaam	lexinfo:partOfSpeech lexinfo:commonNoun
	Gender: mannelijk	lexinfo:gender lexinfo:masculine
	Article: de	lexinfo_anw:articleType
	Meaning class : stofnaam ('substance noun')	:ConceptSchema_ANW-ANS :Concept_SubstanceNoun
Spelling and Flexion		
	Forms	:form_wijn_singular ; :form_wijnen_plural
	Singular form: wijn	lexinfo:number lexinfo:singular
	Singular hyphenation: wijn	lexinfo_anw :hyphenationForm
	Plural form: wijnen	lexinfo:number lexinfo:plural; ontolex:writtenRep
	Plural hyphenation: wijnen	lexinfo_anw:hyphenationForm
Pronunciation		
	Number of syllables: 1	lexinfo_anw:syllable_nb
	Phonetic transcription: *w ɛɪ n	ontolex:phoneticRep
Morphology		
	Type: ongeleed ('simplex')	We have no mapping for this as the ontolex class "Word" is disjoint with the class "MultiWordExpression" and therefore has as instances only "non-compound" words
Usage information	Frequency: 6970	lexinfo_anw:corpusFreq

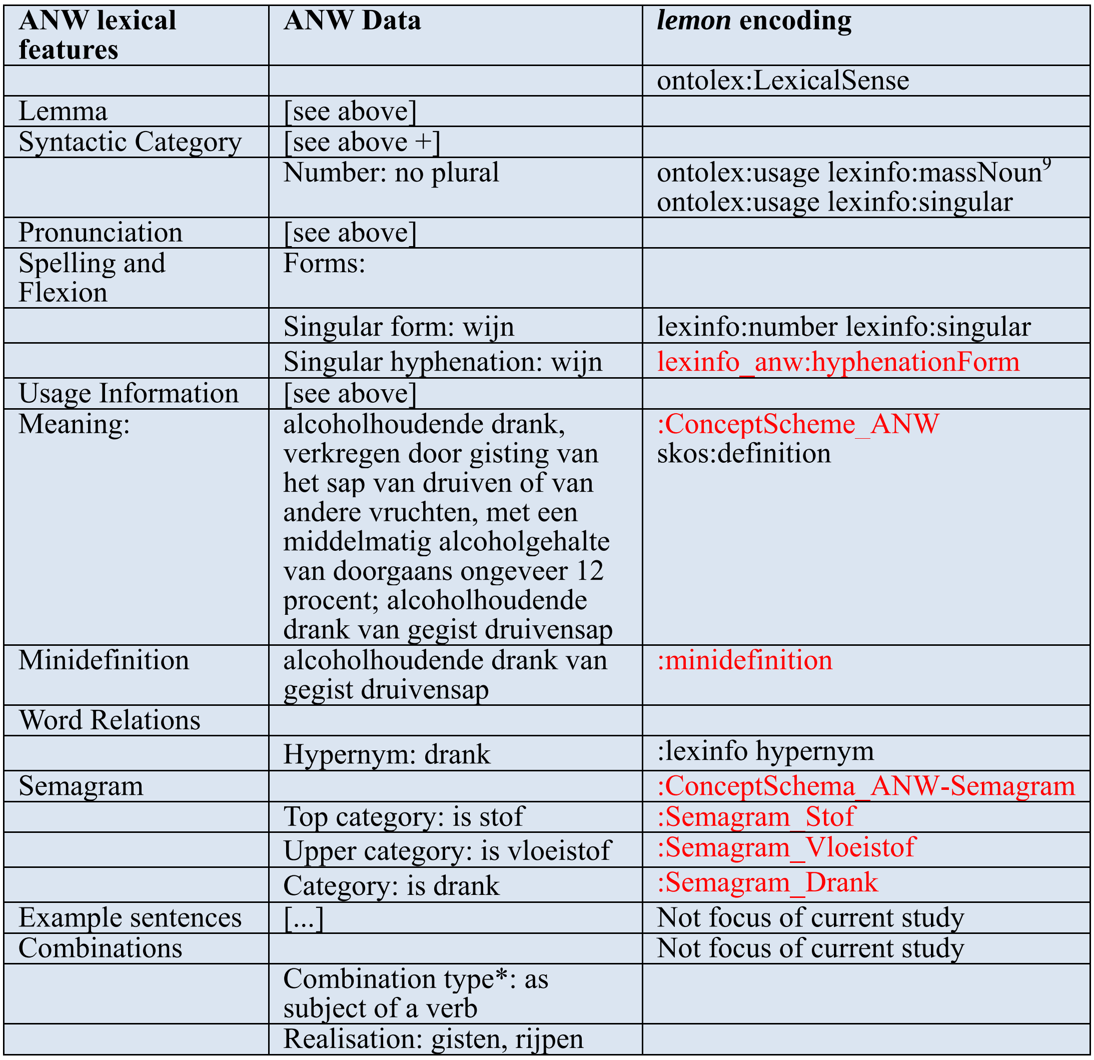
Table 1: Details of the ANW morphosyntactic features for the entry *wijn*

Table 1 shows that most of the morphosyntactic information encoded in the ANW can be coded in *lemon* using the ontolex module and the associated LexInfo vocabulary. Only a few extensions were introduced; for instance, the number of syllables of a word. The encoding of this information is currently not foreseen in LexInfo. However, we feel that this property may also be useful to other lexical resources, therefore we added `lexinfo_anw:syllable_nb`. The same applies to the features hyphenation, frequency and (morphological) type, which do not seem to be language-specific.

An example of a necessary extension that seems to be specific to Dutch, is the feature `lexinfo_anw:articleType,` which contains information on the type of definite article that is required by the nominal lexical entry. This information is encoded in the ANW because in Dutch it is important to know with which definite article a noun can be used. Dutch has two definite articles; some nouns can only be used with the definite article *de*, some can only be used with the definite article *het*, some cannot have a definite article, and some can be used with either definite articles. In some instances, where both articles are possible, there is a preference for either *de* or *het*. We were unsure how to encode this preference information in *lemon*. This issue also applies to labels which mark that a word or meaning is *mostly* used in singular (or in plural) or in a particular language variety or region, etc.

2.2 Encoding of semantic information

Table 2 shows the information structure for the main sense of the lexical entry *wijn*, the sense of an 'alcoholic drink of fermented grape juice'.



ANW lexical features	**ANW Data**	***lemon* encoding**
		ontolex:LexicalSense
Lemma	[see above]	
Syntactic Category	[see above +]	
	Number: no plural	ontolex:usage lexinfo:massNoun[9] ontolex:usage lexinfo:singular
Pronunciation	[see above]	
Spelling and Flexion	Forms:	
	Singular form: wijn	lexinfo:number lexinfo:singular
	Singular hyphenation: wijn	lexinfo_anw:hyphenationForm
Usage Information	[see above]	
Meaning:	alcoholhoudende drank, verkregen door gisting van het sap van druiven of van andere vruchten, met een middelmatig alcoholgehalte van doorgaans ongeveer 12 procent; alcoholhoudende drank van gegist druivensap	:ConceptScheme_ANW skos:definition
Minidefinition	alcoholhoudende drank van gegist druivensap	:minidefinition
Word Relations		
	Hypernym: drank	:lexinfo hypernym
Semagram		:ConceptSchema_ANW-Semagram
	Top category: is stof	:Semagram_Stof
	Upper category: is vloeistof	:Semagram_Vloeistof
	Category: is drank	:Semagram_Drank
Example sentences	[...]	Not focus of current study
Combinations		Not focus of current study
	Combination type*: as subject of a verb	
	Realisation: gisten, rijpen	

[9] Here, we use the LexInfo element `massNoun`, since such a noun is typically uncountable. But we could also introduce a new element `uncountable`, to be more precise and explicit on this feature.

	Example sentences: […]	
Fixed Expressions		Not focus of current study
	Form*: nieuwe wijn in oude zakken (with definition and example sentences)	
Proverbs		Not focus of current study
	Form*: Wijn op bier is plezier en bier op wijn is venijn (definition and example sentences)	
	Form variant: Wijn na bier is plezier en bier na wijn is venijn; ... (including meaning description)	
Word family		Not focus of current study
	Right-headed compounds: abdijwijn; alsemwijn; ...	
	Left-headed compounds: wijnaanbod; wijnacademie; …	
	Derivational compounds: wijnkleurig; wijnmakerij; ...	

Table 2: Details of information for the main sense of the ANW entry *wijn,* sense 1.0

As can be seen in Table 2, the ANW contains semantic information about the lemma in various information categories within the entry, i.e., within the definitions, within the semagrams (an innovative feature of the ANW, described below in Section 2.2.3) and for nouns also in the so-called meaning classes.

2.2.1 Definitions

As any traditional monolingual dictionary, the ANW contains definitions that explain the meaning of the entry. In addition, the ANW provides mini definitions, i.e., short definitions that are used in sense menus to give the user a quick impression of the different senses of a word.

2.2.2 ANS Meaning classes

For nouns, the ANW also classifies the different senses of an entry in so-called meaning classes, a semantic classification of nouns which is based on the *Algemene Nederlandse Spraakkunst (ANS;* Haeseryn et al., 1997*).*

On the basis of Table 3, the following values are distinguished in the ANW: human nouns, animal nouns, object nouns, substance nouns, collective nouns, abstract nouns, proper nouns and plant nouns (an additional value in the ANW). The advantage of having these meaning classes is that it enables lexicographers to provide a global labelling for the sense distinctions. More precise sense information is given in the semagrams in the ANW.

Nouns			**common**	**proper**
concrete	**individual** voorwerpsnamen	**human nouns** persoonsnamen **animal nouns** diernamen **object nouns** zaaknamen	*man ‘man’, meisje ‘girl’, huis ‘house’*	*Jan, Minou, Amsterdam*
	substance stofnamen		*water ‘water’, bier ‘beer’, goud ‘gold’*	
	collective verzamelnamen		*vee ‘cattle’, kroost ‘offspring’, gebergte ‘mountains’*	*Alpen ‘Alps’, Antillen ‘Antilles’*
Abstract			*maand ‘month’, voetbalclub ‘football club’, goedheid ‘kindness’*	*april ‘April’, Vitesse, romantiek ‘romantics’*

Table 3: Semantic classification of Nouns according to the *Algemene Nederlandse Spraakkunst*

2.2.3 Semagrams

Semagrams are an innovative feature of the ANW, which were introduced by Moerdijk (2008), the first editor-in-chief of the ANW. A semagram is the representation of knowledge associated with a word in a frame of ‘slots’ and ‘fillers’. ‘Slots’ are conceptual structure elements which characterise the properties and relations of the semantic class of a word (e.g. COLOUR, SMELL, TASTE, COMPOSITION, INGREDIENTS, PREPARATION for the class of beverages). On the basis of these slots specific data are stored (‘fillers’) for the word in question.

The ANW adopted its own method for defining the semantic classes and the corresponding frames, as it wanted a classification geared towards lexicographic description and based as far as possible on linguistic foundations rather than on a division of words over various social domains. In addition, it wanted a classification which was relatively transparent such that it could also be used in the dictionary’s search function going from content to form. The need to include semagrams in addition to definitions in dictionary entries stems in the first instance from the consideration that definitions alone cannot explain meaning. There is often a lot more semantically relevant knowledge associated with a word than can be shown in a definition. Figure 2 shows the semagram for *wijn* (‘wine’), translated into English for the purpose of this paper.[10] At the moment, only the classification information is

[10] For more information on semagrams, see Moerdijk (2008); Tiberius and Schooheim 2015).

encoded in *lemon.* However, the ontolex model can also be used to encode all additional semantic information, taking advantage of the linkage to the SKOS[11] vocabulary, as can be seen in Figure 1. The work to be done here consists of mapping the ANW semagram into the SKOS structure and then to link the whole SKOS construct to the lexical entry by means of the property `isEvokedBy` and to the corresponding sense of *wijn* with the property `isLexicalizedSenseOf`. The advantage of this approach is that all information from "both" sides of the properties are available using the same representation languages.

Wine: beverage; liquid; substance

- [**Smell**] has depending on the developed aroma bouquet, the odour of earth, red fruit, white flowers, forest scents etc.
- [**Colour**] is mainly red, rose, transparent colourless or yellowish
- [**Taste**] is mildly acidic in the case of red or dry white wine but can depending on the grape variety and fermentation also be semi-sweet, semi-dry or sweet
- [**Transparency**] is generally clear
- [**Ingredient**] is a brew based on fermented juice of fruit, especially of grapes, and contains alcohol, acids, unfermented residual sugar and tannin
- [**Function**] serves to enjoy gastronomically, whether or not during a meal, or is to be drunk for pleasure
- [**Preparation**] is prepared by pressing fruit and allowing the juice to ferment
- [**Raw materials**] is made from the juice of grapes or other fruits
- [**Place of Origin**] is produced worldwide in areas with sufficient sunshine for ripening grapes or other fruit
- [**Container**] is in a bottle, carafe, jar or pack, or is being drained from a barrel
- [**Age**] can be young or old, if suitable as a storage wine
- [**Temperature**] is being drunk cold, cool, at room temperature or warm depending on the type [**Property**] usually has a moderate alcohol percentage, often around 12 percent
- [**Mode of use**] is drunk from a goblet or cup
- [**Working**] can make someone happy, rosy or drunk
- [**Occasion**] is being drunk at meals and during meetings with a certain atmosphere such as parties, ceremonies, a celebration, cosy gathering etc.

Figure 2: Semagram for the lemma *wijn* in the ANW

We have chosen to model the semantic information in the definitions, the semagrams and the meaning classes in the ANW into three SKOS concept sets, i.e.:

:ConceptScheme_ANW (for the definitions)

:ConceptScheme_ANW-ANS (for the ANS meaning classes)

:ConceptSchema_ANW-Semagram (for the semagram)

[11] SKOS stands for "Simple Knowledge Organization System". See also https://www.w3.org/2004/02/skos/ for more details.

In addition, some entries also contain domain information. For instance, the sixth and seventh senses of the entry *kat* 'cat' are marked as belonging to the domain of military history. To model the domain information, we propose to use `dct:subject` from the Dublin Core[12] vocabulary.

On the basis of the above information, the semantic information for the ANW entry for *wijn* is modelled as a skos:Concept which has five lexicalised senses: the main sense and four subsenses. This concept is evoked by the lexical entry for wijn, i.e., `lex_wijn_182155`[13], and the lexical entry for *wijnfles*, i.e., `lex_wijnfles_182210`.

"wijn" lexical entry in *lemon*

```
:Concept_325624
  rdf:type skos:Concept ;
  rdf:type ontolex:LexicalConcept ;
  rdfs:comment "Kernbetekennis for lex_wijn_182155" ;
  skos:inScheme :ConceptScheme_ANW ;
  skos:topConceptOf :ConceptScheme_ANW ;
  ontolex:isEvokedBy :lex_wijn_182155 ;
  ontolex:isEvokedBy :lex_wijnfles_182210 ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.0> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.1> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.2> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.3> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.4> ;
.

:lex_wijn_182155
  rdf:type ontolex:Word ;
  lexinfo_anw:articleType "\"de\"" ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_wijn_singular ;
  ontolex:otherForm :form_wijnen_plural ;
  ontolex:sense <http://tutorial-topbraid.com/anw-entry#sense_wijn1.0> ;
.
```

"form" information for the lexical entry "wijn" in *lemon*

```
:form_wijn_singular
  rdf:type ontolex:Form ;
  <http://lemon-model.net/lexinfo_anw:hyphenationForm> "\"wijn\"" ;
  <http://lemon-model.net/lexinfo_anw:syllable_nb> 1 ;
  dct:language <http://www.lexvo.org/page/iso639-3/nld> ;
  lexinfo:number lexinfo:singular ;
  ontolex:phoneticRep "* w ɛɪ n"@ nl-ReadSpeaker-fonipa ;
  ontolex:writtenRep "wijn"@nl ;
.
:form_wijnen_plural
```

[12] See http://dublincore.org/ for more details

[13] The number refers to the PID of the ANW entry. ANW entries have a PID at the entry level and at the sense level.

```
  rdf:type ontolex:Form ;
  <http://lemon-model.net/lexinfo_anw:hyphenationForm> "\"wij.nen\"" ;
  <http://lemon-model.net/lexinfo_anw:syllable_nb> 2 ;
  dct:language <http://www.lexvo.org/page/iso639-3/nld> ;
  lexinfo:number lexinfo:plural ;
  ontolex:writtenRep "wijnen"@nl ;
.
```

main sense information associated to the lexical entry "wijn" in *lemon*

```
<http://tutorial-topbraid.com/anw#sense_wijn1.0>
  rdf:type ontolex:LexicalSense ;
  skos:definition "alcoholhoudende drank, verkregen door gisting van het sap van
druiven of van andere vruchten, met een middelmatig alcoholgehalte van doorgaans
ongeveer 12 procent; alcoholhoudende drank van gegist druivensap" ;
  ontolex:isLexicalizedSenseOf :Concept_325624 ;
  ontolex:isLexicalizedSenseOf :Concept_Stofnaam ;
  ontolex:isLexicalizedSenseOf :Concept_mass ;
  ontolex:isLexicalizedSenseOf :Semagram_drank ;
  ontolex:isSenseOf :lex_wijn_182155 ;
  ontolex:reference <https://www.wikidata.org/wiki/Q282> ;
  ontolex:usage lexinfo:massNoun ;
  ontolex:usage lexinfo:singular ;
.
```

subssenses originally associated to the entry "wijn", here in the *lemon* encoding

```
<http://tutorial-topbraid.com/anw#sense_wijn1.1>
  rdf:type ontolex:LexicalSense ;
  skos:definition "wijnsoort of wijnmerk" ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
.
<http://tutorial-topbraid.com/anw#sense_wijn1.2>
  rdf:type ontolex:LexicalSense ;
  skos:definition "druiven gekweekt als gewas voor de wijnproductie; wijndruiven als
gewas" ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
.
<http://tutorial-topbraid.com/anw#sense_wijn1.3>
  rdf:type ontolex:LexicalSense ;
  lemon:broader <http://tutorial-topbraid.com/anw#sense_fles1.0> ;
  rdfs:td_is_container_of <http://tutorial-topbraid.com/anw#sense_wijn1.0> ;
  skos:definition "fles wijn" ;
  ontolex:isLexicalizedSenseOf :Concept_325624 ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
  ontolex:reference <https://www.wikidata.org/wiki/Q23490> ;
.
<http://tutorial-topbraid.com/anw#sense_wijn1.4>
  rdf:type ontolex:LexicalSense ;
  lexinfo:partMeronym <http://tutorial-topbraid.com/anw-entry#sense_wijn1.0> ;
  skos:definition "portie of hoeveelheid wijn; glas wijn" ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
  ontolex:reference
<https://commons.wikimedia.org/wiki/File:Glass_wine_white_background.jpg> ;
.
```

2.3 Encoding of compounds

To represent ANW compounds in *lemon*, we make use of the decomposition module, which is depicted in Figure 3 below. An important point being that at this stage we consider compounds as an instance of the MultiWordExpression class of ontolex (see the graphical representation of the ontolex module further above).

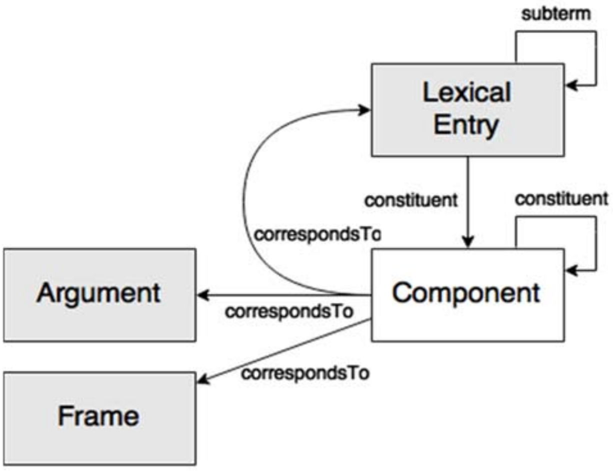


Figure 3: The decomposition module of lemon[14]

In the following *lemon* code below, we can see how the word *wijnfles* ('wine bottle') is decomposed in both its surface form elements (via the property `constituent`) and its compounding lexical entries (via the property `subterm`). The ordering of the elements of the compound is marked with the rdf construct `rdf_1`, etc. The whole compound entry is listed as having the sense `sense_wijn1.3,` which itself is one of the senses for the entry *wijn*. This example shows the potential of *lemon* for sharing and re-using elements of the lexicon across the whole dictionary, and also for linking to other data sources, as every element is encoded internally as a unique resource identifier (URI), including its location on the web.

```
:lex_wijnfles_182210
  rdf:type ontolex:MultiWordExpression ;
  lexinfo_anw:articleType "\"de\"" ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  rdf:_1 :comp_wijn_1 ;
  rdf:_2 :comp_fles_1 ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :comp_fles_1 ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :comp_wijn_1 ;
  <http://www.w3.org/ns/lemon/decomp#subterm>
<http://dictionary_lemon/anw#lex_wijn_182155> ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :lex_fles_18089 ;
  ontolex:sense <http://tutorial-topbraid.com/anw#sense_wijn1.3> ;
.
```

[14] Figure created by John P. McCrae for the W3C Ontolex Community Group.

3. Concluding remarks

In this paper, we have presented a *lemon* model for the morphosyntactic and semantic information in the ANW, a comprehensive scholarly dictionary of Dutch. Encoding the information in *lemon* has a number of advantages:

- **Modularization of the data**

As we could observe especially in the case of the representation of compounds, the *lemon* model implements a strong modular approach to the encoding of lexicon data, and therefore strongly supports the re-use of such elements. This is also true when we look at the internal XML encoding of the ANW, in which for every sense of an entry the whole morphosyntactic information—with some local variations—has to be repeated. This can be avoided in the *lemon* model, as all the different elements of an entry are modularly encoded and interlinked by specific interpretation. There is no redundancy in the graph-based *lemon* model.

- **Linking**

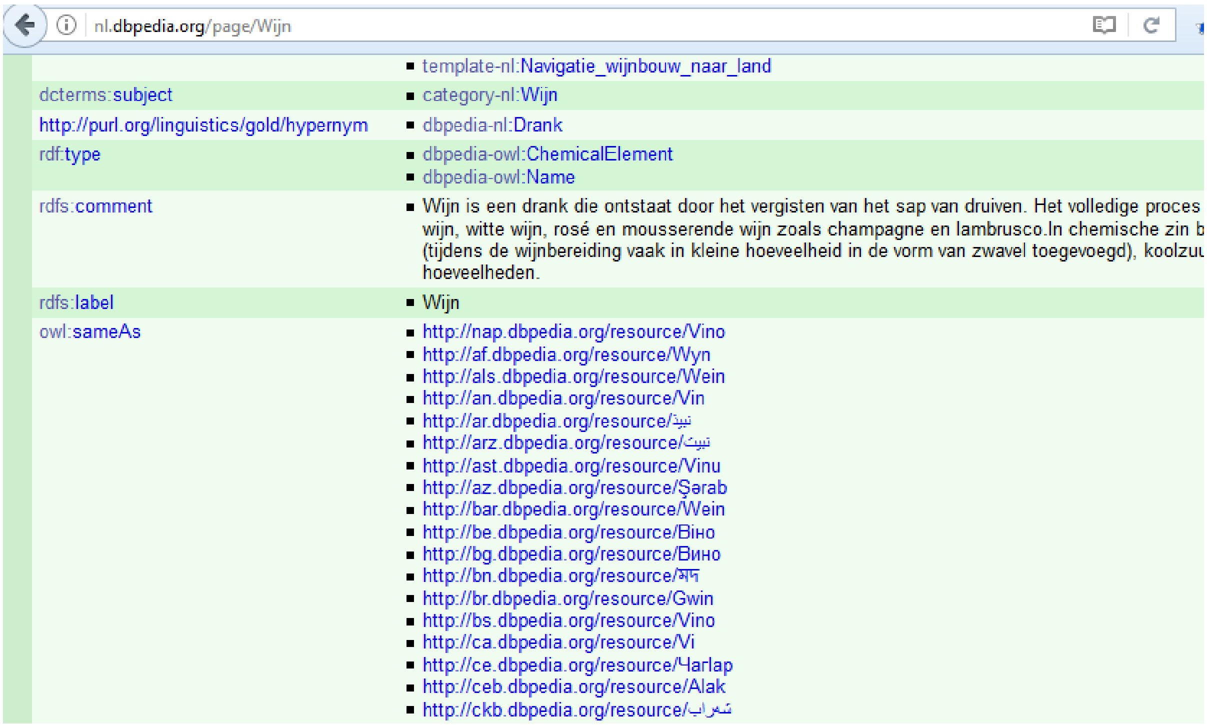
As the lemon model is making use of W3C standards for encoding its elements, linking is the major way to express relations between such elements within one dictionary, but also for external data sources that are encoded as an URI (with a valid location). In the case of the *wijn* entry, we are for example linking the sense 1.0 to a wikidata[15] entry and to a DBpedia[16] entry:

```
<http://tutorial-topbraid.com/anw#sense_wijn1.0>
  rdf:type ontolex:LexicalSense ;
  …
  ontolex:isSenseOf :lex_wijn_182155 ;
  ontolex:reference <https://www.wikidata.org/wiki/Q282> ;
  ontolex:reference http://nl.dbpedia.org/page/Wijn ;
  …
.
```

Accessing then the wikidata or the DBpedia location, one can gain additional information, for example a relevant number of translations of the word *wijn* in this particular sense, as the screenshot of the (partial) page of DBpedia shows in Figure 4.

[15] See https://www.wikidata.org/wiki/Wikidata:Main_Page.

[16] See http://wiki.dbpedia.org/.



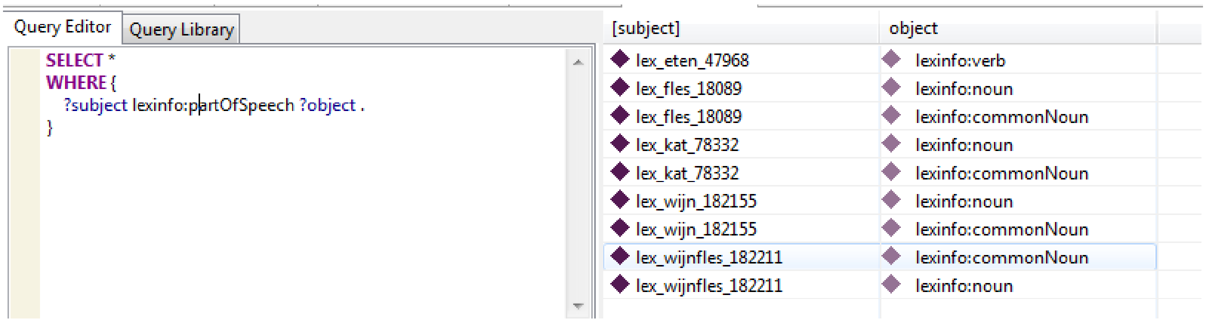
nl.dbpedia.org/page/Wijn

	template-nl:Navigatie_wijnbouw_naar_land
dcterms:subject	category-nl:Wijn
http://purl.org/linguistics/gold/hypernym	dbpedia-nl:Drank
rdf:type	dbpedia-owl:ChemicalElement dbpedia-owl:Name
rdfs:comment	Wijn is een drank die ontstaat door het vergisten van het sap van druiven. Het volledige proces wijn, witte wijn, rosé en mousserende wijn zoals champagne en lambrusco.In chemische zin b (tijdens de wijnbereiding vaak in kleine hoeveelheid in de vorm van zwavel toegevoegd), koolzuu hoeveelheden.
rdfs:label	Wijn
owl:sameAs	http://nap.dbpedia.org/resource/Vino http://af.dbpedia.org/resource/Wyn http://als.dbpedia.org/resource/Wein http://an.dbpedia.org/resource/Vin http://ar.dbpedia.org/resource/نبيذ http://arz.dbpedia.org/resource/نبيت http://ast.dbpedia.org/resource/Vinu http://az.dbpedia.org/resource/Şərab http://bar.dbpedia.org/resource/Wein http://be.dbpedia.org/resource/Віно http://bg.dbpedia.org/resource/Вино http://bn.dbpedia.org/resource/মদ http://br.dbpedia.org/resource/Gwin http://bs.dbpedia.org/resource/Vino http://ca.dbpedia.org/resource/Vi http://ce.dbpedia.org/resource/Чarlap http://ceb.dbpedia.org/resource/Alak http://ckb.dbpedia.org/resource/شەراب

Figure 4: The DBpedia page on ‘wijn’

- **Query and access to the data**

The dictionary data encoded in *lemon* are stored in so-called triple stores and thus can be queried and are accessible by the use of the standardised SPARQL query language[17]. It is worth mentioning here, that SPARQL can also be used for augmenting the original data set. The main point is the fact that the ANW data can, in this way, be made available for processing engines, since it is now in a fully machine-readable format. Below we show an example of a simple query we performed with the TopBraid composer[18]. On the left is the query and on the right the results. In this example, the query asks for all entries that have a part-of-speech, while also querying for information about the part-of-speech.



```
SELECT *
WHERE {
  ?subject lexinfo:partOfSpeech ?object .
}
```

[subject]	object
lex_eten_47968	lexinfo:verb
lex_fles_18089	lexinfo:noun
lex_fles_18089	lexinfo:commonNoun
lex_kat_78332	lexinfo:noun
lex_kat_78332	lexinfo:commonNoun
lex_wijn_182155	lexinfo:noun
lex_wijn_182155	lexinfo:commonNoun
lex_wijnfles_182211	lexinfo:commonNoun
lex_wijnfles_182211	lexinfo:noun

[17] https://www.w3.org/TR/rdf-sparql-query/.

[18] http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition/.

In general, we can state that the ontolex and the decomposition modules of *lemon* could be used as they are, while the modifications needed for being compliant with the richness of the ANW data can be addressed in the context of the LexInfo vocabulary, and our ongoing work is to make sure that the inclusion of those ANW features are either made part of LexInfo, or are made available within a similar ontology.

4. Acknowledgements

The DFKI contribution to this paper was partly supported by the H2020 project QT21 with agreement number 645452.

5. References

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Aguado-de-Cea, G. (2016). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop*, pp. 65-73.

Declerck, Th., Wandl-Vogt, E., Krek, S. & Tiberius, C. (2015). Towards Multilingual eLexicography by Means of Linked (Open) Data. *MSW@ESWC 2015*, pp. 51-58.

Declerck, Th. & Mörth, K. (2016). 'Towards a Sense-based Access to Related Online Lexical Resources'. In: T. Margalitadze & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress, Tbilissi, Georgia*, pp. 660-667.

Khan, F., Bellandi, A., Boschetti, F. & Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In: *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017).*

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., ven den Toorn, M. C. (1997). *Algemene Nederlandse spraakkunst.* Groningen Nijhoff.

El Maarouf, I., Jane Bradbury, J., & Hanks, P. (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In: *Proceedings of the 3rd Workshop on Linked Data in Linguistics,* pp-87-93.

McCrae, J. P., Cimiano, P., Buitelaar, P. & Bordea, G. (2016a). 'Representing Multiword Expressions on the Web with the OntoLex-Lemon model'. In *PARSEME/ENeL workshop on MWE e-lexicons.*

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, Th., de Melo, G., Gracia, K., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A. & Pool, J. (2016b), 'The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud'. In *Proceedings of the 10th Language Resource and Evaluation Conference (LREC).*

McCrae, J. P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, Th., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. &

Wunner, T. (2012). ‘Interchanging lexical resources on the Semantic Web’. In *Language Resources and Evaluation*, 46(6), pp. 701-709.

Moerdijk, F. (2008). 'Frames and Semagrams. Meaning Description in the General Dutch Dictionary'. In: E. Berndal & J. De Cesaris (eds.) *Proceedings of the XIII EURALEX International Congress, Barcelona*, pp. 561-569.

Schoonheim, T. & Tempelaars, R. (2010). 'Dutch Lexicography in Progress, The Algemeen Nederlands Woordenboek (ANW)'. In: A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress. Leeuwarden*, pp. 718-725.

Stolk, S. (2017). OntoLex and Onomasiological Ordering: Supporting Topical Thesauri. In: *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017).*

Tiberius, C. & Schoonheim, T. (2015). Semagrams, Another Way to Capture Lexical Meaning in Dictionaries’. In *Journal of Cognitive Science*, 16(4), pp. 379-400.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/