Clariah-WP3
Interoperability of lexical resources

Interoperability of Lexical Resources

DICOLOD DIacronous COnceptual lexicons through Linked Open Data Documentation

Author:
Isa Maks (isa.maks@vu.nl)

Version:
Draft 0.1

Contents

1 Introduction

Please note that this is a first version of the documentation as DICOLOD is still an ongoing project.

This document describes recent developments in the ongoing project DICOLOD. DICOLOD is a collection of existing lexical resources describing words and word senses of various historical and contemporaneous variants of Dutch. The resources have been compiled in different periods of time and most of them have been digitized in a later phase. All of them are semantic resources having some kind of ontological backbone such as a thesaurus classification an organisation into sets of synonyms, or another kind of semantic typology.

We aim at composing a resource that describes the vocabulary of Dutch ranging from the 16th to the 21st century by bringing these resources together and converting them into RDF triples . Obviously, the completeness depends on the number and size of the resources and the periods covered by them. By converting the resources to rdf triples and linking them on the conceptual level, we aim at an overview on how words and concepts develop over time.

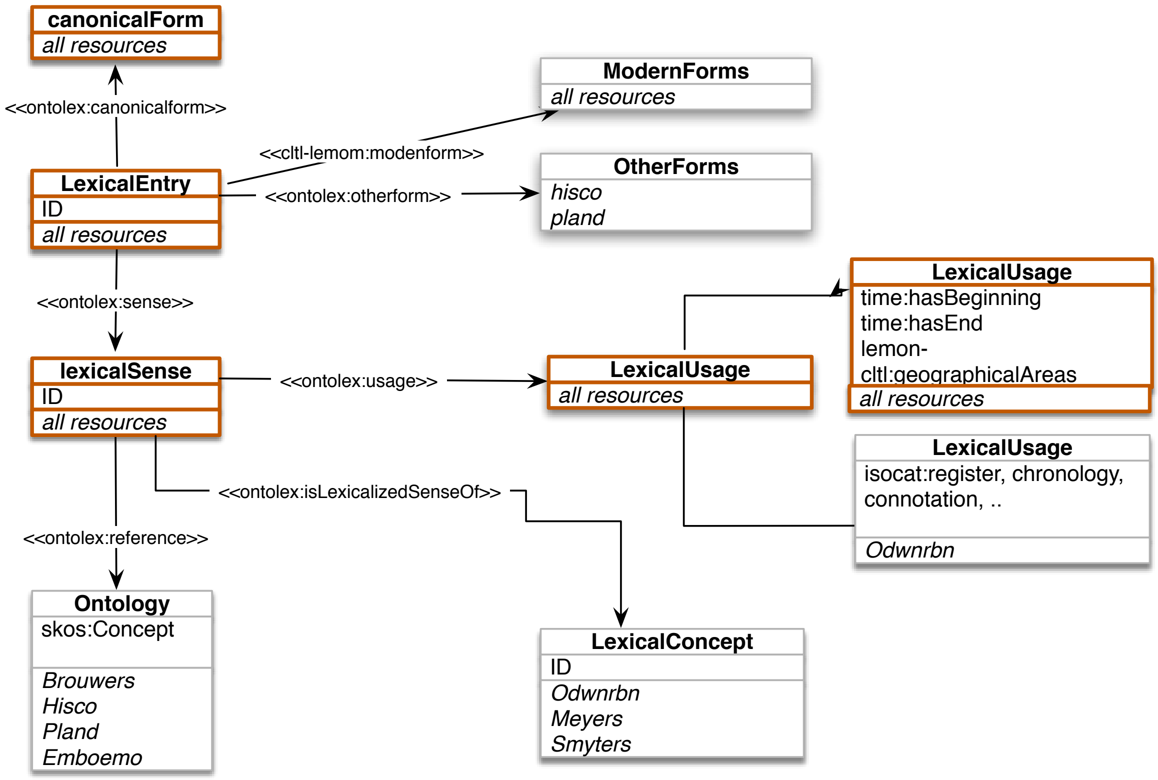
One of the prerequisites of the project is that all included resources are publicly available in order to be able to publish all results of the project as linked open data.

The project consists of the following steps:

- design of a common data model which includes the data of the different resources and which captures the information available in the different resources, enables the linking between them and records the period of time in which the word in a particular meaning is used

- conversion of all the resources to the common data model: the involved lexicons which are mostly in their own proprietary (non-standardized) formats available are converted to the common model.

- automatic and semi-automatic generation of mapping links that interlink the different resources.



LOD model for diachronous concept lexicons (DICOLOD) - preliminary version (Jan 2017)-
e.maks@vu.nl

Figure 1: DICOLOD-lemon model (see section 2.2)

2 Datamodel

The first step consists of the conversion of the data to one common data model that follows the LEMON - Lexicon Model for Ontologies standard (Lemon) in combination with the Lexicon Markup Framework (http://www.lexicalmarkupframewor

2.1 Used standards

2.1.1 Lemon

lemon (Lexicon Model for Ontologies) [Mccrae,] is a model for associating linguistic information with ontologies, in particular Semantic Web ontologies. Lemon separates the lexical layer, that is the words and their morphology and syntactic behaviour, and the semantic layer in the ontology, which describes the meaning of the entry. The model of lemon is based around the object LexicalSense which refers to one meaning of a word and which connects to ontology entitiesFor the description of the linguistic information Lemon makes use of the LMF standard, for the description of the semantic information lemon refers to Skos. WordNets have an own submodel in Lemon which is further explained below

2.1.2 Lmf

the lexical information layer of the resources is modeled according to the LMF (Lexicon Markup Framework) standard. The model of LMF centers around lexicalEntry, corresponding to a word, and lexicalSense representing one meaning of the word. All linguistic information such as part-of-speech, orthographical, morphological, syntactical , etc. information can be modeled in LMF and attached to the LexicalEntry or the LexicalSense.

2.1.3 Skos

- skos:Concept
- skos:prefLabel
- skos:altLabel
- skos:broader

[0] http://www.w3.org/2016/05/ontolex/

- skos:related

2.2 The DICOLOD data model

Figure (1) shows the DICOLOD model with the lemon elements(Ontology, LexicalSense) and relations (ontolex:reference and ontolex:isLexicalizedSenseOf). For DICOLOD we put special attention on the usage element and the element modernForm, both not yet defined in LMF.

An example of the information in RDF-triples can be found below.

- In our model a LexicalEntry always corresponds to one lexicalSense, which implies that all information provided with the lexicalEntry applies to this particular sense.

- The LEMON links between the lexical sense and the ontology is called 'reference' and 'isReferenceOf'. This is a equivalence link which means that the concept in the ontology must be the same concept as the sense it is referring to. Therefore we create first a dummy-concept (cf. emboemo:concept-22) which can be linked to the ontology.

- we added modernForm to all data sets which will help to interlink the data. The forms are retrieved by using the INL lexicon service (REF)

- the ontolex:usage component is important for the representation of the historical data. We defined the component specifically for this project as it is not part of Lemon or LMF. Usage information , in our case, refers to the period in which the LU is used, and the geographical space in which it is used. In the following example the usage component 'emboemo:usage-220' is created to capture this information.

```
\label{entry-vb}
emboemo:entry-22 a ontolex:LexicalEntry .
emboemo:entry-22 ontolex:canonicalForm "pyn" .
emboemo:entry-22 rdfs:label "pyn"@nl .
emboemo:entry-22 lexinfo:partOfSpeech lexinfo:noun .
emboemo:entry-22 lemon-cltl:modernForm "pijn"  .
emboemo:entry-22 ontolex:sense emboemo:sense-22 .
emboemo:sense-22 a ontolex:LexicalSense .
emboemo:sense-22 ontolex:reference emboemo:concept-22 .
emboemo:sense-22 ontolex:usage emboemo:usage-220 .
emboemo:usage-220 a ontolex:LexicalUsage .
emboemo:usage-220 time:hasBeginning time:Instant-1600 .
time:Instant-1600 time:year "1600"^^xsd:gYear .
emboemo:usage-220 time:hasEnd time:Instant-1830 .
```

```
time:Instant-1830 time:year "1830"^^xsd:gYear .
emboemo:usage-220 lemon-cltl:geographicalArea "Netherlands" .
emboemo:usage-220 lemon-cltl:geographicalArea "Belgium" .
```

2.3 Namespaces

In all data sets the following namespaces are used

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> . (lemon and lmf)
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .  (part-of-speech and other linguistic information
@prefix wnaffect: <http://www.gsi.dit.upm.es/ontologies/wnaffect/ns#> .  (classification of emotions linked to wordnet)
@prefix skos: <http://www.w3.org/2004/02/skos/core#> . (description of ontological information)
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .
@prefix xsd:  <http://www.w3.org/2000/10/XMLSchema#> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix isocat: <http://www.isocat.org/datcat/> . (wordnet relations)
@prefix lemon-cltl: <http://cltl.nl/clariah-lemon.placeholder/> . (specific usage information needed for the diacronous
```

Additionally, for each specific lexicon an own namespace is defined:

```
@prefix odwnrbn: <http://cltl.nl/odwnrbn.placeholder/> .
@prefix emboemo: <http://cltl.nl/emboemo.placeholder/> .
@prefix hisco: <http://cltl.nl/hisco.placeholder/> .
@prefix brouwers: <http://cltl.nl/brouwers.placeholder/> .
@prefix smyters: <http://cltl.nl/smyters.placeholder/> .
@prefix pland: <http://cltl.nl/pland.placeholder/> .
```

2.4 Provenance and versioning

still to do

3 Data sets

3.0.1 Emboemo

Emboemo is the lexicon that is constructed of the HEEM project [1] [?] and contains words denoting emotions as found in 16-18th century drama texts. It is based in corpus annotations and as such it contains non-lemmatized forms with part-of-speech labels. The part-of-speech labels may be incorrect because of tagging errors. The words are categorised in 36 emotion classes.

[1] bla

The relation between the ontolex:lexicalSense and the emboemo:concept is that of a SKOS:broader.

lexical unit triples

```
emboemo:entry-63 a ontolex:LexicalEntry .
emboemo:entry-63 ontolex:canonicalForm "toornicheijt" .
emboemo:entry-63 rdfs:label "toornicheijt"@nl .
emboemo:entry-63 lexinfo:partOfSpeech lexinfo:noun .
emboemo:entry-63 ontolex:sense emboemo:sense-63 .
emboemo:sense-63 a ontolex:LexicalSense .
emboemo:sense-63 ontolex:reference emboemo:concept-63 .
emboemo:sense-63 ontolex:usage emboemo:usage-630 .
emboemo:usage-630 a ontolex:LexicalUsage .
emboemo:usage-630 time:hasBeginning time:Instant-1600 .
time:Instant-1600 time:year "1600"^^xsd:gYear .
emboemo:usage-630 time:hasEnd time:Instant-1830 .
time:Instant-1830 time:year "1830"^^xsd:gYear .


emboemo:concept-63 a skos:Concept .
emboemo:concept-63 ontolex:isReferenceOf emboemo:sense-63 .
emboemo:concept-63 skos:prefLabel "toornicheijt"@nl .
emboemo:concept-63 skos:broader emboemo:concept-Woede .
```

emotion classifiction

```
emboemo:concept-Woede a skos:Concept .
emboemo:concept-Woede skos:prefLabel "woede"@nl .
emboemo:concept-Woede skos:broader emboemo:Emotie .
```

3.0.2 Brouwers

Brouwers is a thesaurus first published in 1930 and re-published in 1978 which is recently digitized by converting the pdf book scan to a csv format. Not all information is correctly processed and especially the classification shows inconsistencies which deviate from how it is originally presented in the book. Therefore, the data are preprocessed in order to clean them (unnecessary punctuation and numbering is removed), and in order to remedy classification errors by replacing them with the original classes as found in the book.[2] (REF Appendix for overview replacements). The pre-processing is still an ongoing process as classification errors become apparent only when

[2] For example, 'Tweezaadlobbigen_Dicotylodonae_II' (a type of plants) is directly classified under the high level class 'Stoffelijke wezens' whereas the intermediate class 'Plantenrijk' not appeared in the csv file. We re-inserted 'Planten' as intermediate class.

the data is inspected more carefully.

As the following example shows the link between the lexicon and the ontology is done by a reference link to a dummy skos concept (concept-132806). Brouwers is a thesaurus with an ontological organisation structure where words are classified into topical classes which are part of a hierarchical organisation. The words that belong to a class may have any relation (synonym, antonym, holonym, part-of, etc.) with the other members of the class. Hence we defined the relations between the classes as skos:broader relations, whereas the relations between the reference concept and the lowest class of the ontology is a skos:related relation.

lexical sense triples

```
brouwers:entry-132806 a ontolex:LexicalEntry .
brouwers:entry-132806 ontolex:canonicalForm "verdriet" .
brouwers:entry-132806 rdfs:label "verdriet"@nl .
brouwers:entry-132806 lexinfo:partOfSpeech lexinfo:noun .
brouwers:entry-132806 ontolex:sense brouwers:sense-132806 .
brouwers:sense-132806 a ontolex:LexicalSense .
brouwers:sense-132806 ontolex:reference brouwers:concept-132806 .
brouwers:sense-132806 ontolex:usage brouwers:usage-1328060 .
brouwers:usage-1328060 a ontolex:LexicalUsage .
brouwers:usage-1328060 time:hasBeginning time:Instant-1850 .
time:Instant-1850 time:year "1850"^^xsd:gYear .
brouwers:usage-1328060 time:hasEnd time:Instant-1975 .
time:Instant-1975 time:year "1975"^^xsd:gYear .
brouwers:usage-1328060 lemon-cltl:geographicalArea "Netherlands" .
brouwers:usage-1328060 lemon-cltl:geographicalArea "Belgium" .
```

classification triples

```
brouwers:concept-132806 a skos:Concept .
%brouwers:concept-132806 skos:note "zielenlijden*;Lijden;Vreugde_droefheid;Gevoelens;" .
brouwers:concept-132806 skos:prefLabel "verdriet"@nl .
brouwers:concept-132806 ontolex:isReferenceOf brouwers:sense-132806 .
brouwers:concept-132806 skos:related brouwers:concept-zielenlijden* .

brouwers:concept-zielenlijden* a skos:Concept .
brouwers:concept-zielenlijden* skos:prefLabel "zielenlijden"@nl .
brouwers:concept-zielenlijden* skos:related brouwers:concept-Lijden .

brouwers:concept-Lijden a skos:Concept .
brouwers:concept-Lijden skos:prefLabel "lijden"@nl .
brouwers:concept-Lijden skos:related brouwers:concept-Vreugde_droefheid .

brouwers:concept-Vreugde_droefheid a skos:Concept .
brouwers:concept-Vreugde_droefheid skos:prefLabel "vreugde droefheid"@nl .
brouwers:concept-Vreugde_droefheid skos:related brouwers:concept-Gevoelens .
```

3.0.3 OdwnRbn

In LEMON, WordNets words are regarded as lemon lexical entries and the word senses correspond to lemons lexical senses ([Mccrae,]). WordNet's synsets are regarded as ontological references for which a new type lexicalConcept as a subclass of Concept in SKOS is introduced. The reference of the lexical senses are these lexicalConcepts. In this way the nature of synsets is captured without ontologizing the semantic network. Likewise, relations such as hypernymy, meronymy etc. are introduced as new properties instead of relating them to existing ontological properties such as OWLs subClassOf. Lemon has specific classes to describe wordnet and wordnet-like resources where the lexical units are organised in sets of synonyms.

pland:entry-125708 ontolex:otherForm ”blaauwe knpies”

- LEMON : A synset is called a *lexicalConcept*
- LEMON : Links between lexicalSense and the lexicalConcept are called isLexicalizedSenseOf and LexicalSense
- wordnet relations between synsets are defined in the Isocat Data Category Registry (isocat.org).
- We translated the hyponym hierarchy in a skos hierarchy making the synsets skos:Concepts (which is not in accordance with LEMON) and translating wordnet relations into SKOS:broader relations. (explain why)
  - hyponym (is translated in) narrower
  - hyperonym (is translated in) broader
  - all other wordnet relations (is translated in) related
- If relevant a reference to the external classication WordNetAffect is given.

The data are converted from Open Dutch WordNet[3] REF which includes synsets and lexical units linked to the members of the individual synsets. All data from the wordnet part are included, but we only included only those parts of the lexical unit information which is needed for compiling

[3] https://github.com/cltl/OpenDutchWordnet

DICOLOD. We included semantic and usage information and part-of-speech, but for instance detailed syntactic information is left out. If needed, extra information can be converted in later stage and added.

lexical sense triples

```
odwnrbn:entry-verdriet-n-2 a ontolex:LexicalEntry .
odwnrbn:entry-verdriet-n-2 ontolex:canonicalForm "verdriet" .
odwnrbn:entry-verdriet-n-2 rdfs:label "verdriet"@nl .
odwnrbn:entry-verdriet-n-2 lexinfo:partOfSpeech lexinfo:noun .
odwnrbn:entry-verdriet-n-2 ontolex:sense odwnrbn:sense-verdriet-n-2 .
odwnrbn:sense-verdriet-n-2 a ontolex:LexicalSense .
odwnrbn:sense-verdriet-n-2 ontolex:isLexicalizedSenseOf odwnrbn:synset-eng-30-07534430-n .
odwnrbn:sense-verdriet-n-2 ontolex:usage odwnrbn:usage-verdriet-n-20 .
odwnrbn:usage-verdriet-n-20 a ontolex:LexicalUsage .
odwnrbn:usage-verdriet-n-20 time:hasBeginning time:Instant-1950 .
time:Instant-1950 time:year "1950"^^xsd:gYear .
odwnrbn:usage-verdriet-n-20 time:hasEnd time:Instant-2010 .
time:Instant-2010 time:year "2010"^^xsd:gYear .
odwnrbn:usage-verdriet-n-20 lemon-cltl:geographicalArea "Netherlands" .
odwnrbn:usage-verdriet-n-20 lemon-cltl:geographicalArea "Belgium" .
```

synset triples

```
odwnrbn:synset-eng-30-07534430-n a skos:Concept .
odwnrbn:synset-eng-30-07534430-n skos:definition "an emotion of sadness associated with loss"@en
.
odwnrbn:synset-eng-30-07534430-n skos:broader wnaffect:sadness .
odwnrbn:synset-eng-30-07534430-n ontolex:lexicalizedSense odwnrbn:sense-pijn-n-3 .
odwnrbn:synset-eng-30-07534430-n ontolex:lexicalizedSense odwnrbn:sense-treurigheid-n-4 .
odwnrbn:synset-eng-30-07534430-n ontolex:lexicalizedSense odwnrbn:sense-treurnis-n-3 .
odwnrbn:synset-eng-30-07534430-n ontolex:lexicalizedSense odwnrbn:sense-triestheid-n-3 .
odwnrbn:synset-eng-30-07534430-n ontolex:lexicalizedSense odwnrbn:sense-verdriet-n-2 .
odwnrbn:synset-eng-30-07534430-n ontolex:lexicalizedSense odwnrbn:sense-wee-n-4 .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5824#has_hyperonym odwnrbn:synset-eng-30-07532440-n
.
odwnrbn:synset-eng-30-07534430-n skos:broader odwnrbn:synset-eng-30-07532440-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5881#xpos_near_synonym odwnrbn:synset-eng-30-01796582-v
.
odwnrbn:synset-eng-30-07534430-n skos:related odwnrbn:synset-eng-30-01796582-v .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5881#xpos_near_synonym odwnrbn:synset-eng-30-01797051-v
.
odwnrbn:synset-eng-30-07534430-n skos:related odwnrbn:synset-eng-30-01797051-v .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-eng-30-07535010-n
.
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-eng-30-07535010-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-odwn-10-107767909-n
.
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-odwn-10-107767909-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-eng-30-07538272-n
.
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-eng-30-07538272-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-odwn-10-108045568-n
.
```

```
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-odwn-10-108045568-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-odwn-10-105816686-n
.
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-odwn-10-105816686-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-eng-30-07534847-n
.
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-eng-30-07534847-n .
odwnrbn:synset-eng-30-07534430-n isocat:DC-5823#has_hyponym odwnrbn:synset-eng-30-07535209-n
.
odwnrbn:synset-eng-30-07534430-n skos:narrower odwnrbn:synset-eng-30-07535209-n .
```

3.1 Pland

3.2 Smyters

3.3 Meyers

3.4 Hisco

4 Links between the resources

Links between the different datasets are generated in a automatic or manual way. Link mappings are provided with skos:exactMatch or owl:sameAs predicates and they are accompanied with confidence scores and provenance information.

Example:

```
emboemo:concept-22 skos:exactMatch brouwers:concept-132806

emboemo:concept-22 a skos:Concept .
emboemo:concept-22 ontolex:isReferenceOf emboemo:sense-22 .
emboemo:concept-22 skos:prefLabel "pyn"@nl .
emboemo:concept-22 skos:broader emboemo:concept-Verdriet .

brouwers:concept-132806 a skos:Concept .
brouwers:concept-132806 skos:note "zielenlijden*;Lijden;Vreugde_droefheid;Gevoelens;" .
brouwers:concept-132806 skos:prefLabel "verdriet"@nl .
brouwers:concept-132806 ontolex:isReferenceOf brouwers:sense-132806 .
brouwers:concept-132806 skos:related brouwers:concept-zielenlijden .
```

5 Conversion tools

5.1 Conversion of lexicon with 'flat' hierarchy to LOD

to do

5.2 Conversion of ODWN to LOD

to do

References

[Mccrae,] Mccrae, J. Publishing and linking wordnet using lemon and rdf.

6 Appendix A: replacements in brouwers

```
cp brouwers-inter.xml brouwers-interr.xml


#wijziging
echo Belastingen

sed   -i.bak ’s/Belastingen;In_het_algemeen;Gevoelens/Belastingen;In_het_algemeen;


#wijziging t.o.v. Brouwers: kunsten niet meer onder Gevoelens maar als aparte (nie
echo Kunst
sed   -i.bak ’s/Muziek;Schoonheidsgevoel;Gevoelens/Muziek;Kunst/g’ brouwers-interr
sed   -i.bak ’s/Muziekinstrumenten;Schoonheidsgevoel;Gevoelens/Muziekinstrumenten;
sed   -i.bak ’s/kunst;Schoonheidsgevoel;Gevoelens/kunst;Kunst/g’ brouwers-interr.x
sed   -i.bak ’s/Schoonheidsgevoel;Gevoelens/Schoonheidsgevoel/g’ brouwers-interr.x

#wijziging t.o.v. Brouwers: kunsten niet meer onder Gevoelens maar als aparte (nie
echo Sport
sed   -i.bak ’s/Balspel;Vreugde_droefheid;Gevoelens/Balspel;Sport;Sport_en_spel/g’
sed   -i.bak ’s/Schaatssport;Vreugde_droefheid;Gevoelens/Schaatssport;Sport;Sport_
sed   -i.bak ’s/Gymnastiek;Vreugde_droefheid;Gevoelens/Gymnastiek;Sport;Sport_en_s
sed   -i.bak ’s/Sportwedstrijd;Vreugde_droefheid;Gevoelens/Sportwedstrijd;Sport;Sp
sed   -i.bak ’s/Worstelwedstrijd;Vreugde_droefheid;Gevoelens/Worstelwedstrijd;Spor
sed   -i.bak ’s/Biljartspel;Vreugde_droefheid;Gevoelens/Biljartspel;Spel;Sport_en_
sed   -i.bak ’s/Volksspelen;Vreugde_droefheid;Gevoelens/Volksspelen;Spel;Sport_en_
sed   -i.bak ’s/Gezelschapsspelen;Vreugde_droefheid;Gevoelens/Gezelschapsspelen;Sp
sed   -i.bak ’s/Kaartspel;Vreugde_droefheid;Gevoelens/Kaartspel;Spel;Sport_en_spel
sed   -i.bak ’s/Kinderspelen;Vreugde_droefheid;Gevoelens/Kinderspelen;Spel;Sport_e
sed   -i.bak ’s/Dans;Vreugde_droefheid;Gevoelens/Dans;Sport_en_spel/g’ brouwers-in
sed   -i.bak ’s/Spel_en_sport;Vreugde_droefheid;Gevoelens/Sport_en_spel/g’ brouwer

#correctie invoer : categorie Planten;Plantenrijk tussenvoegen zoals in Brouwers b
echo Plant
```

```
sed   -i.bak ’s/Tweezaadlobbigen_Dicotylodonae_II/Tweezaadlobbigen_Dicotylodonae_I
sed   -i.bak ’s/Tweezaadlobbigen_Dicotylodonae_I/Tweezaadlobbigen_Dicotylodonae_I;
sed   -i.bak ’s/Eenzaadlobbigen_Monocotyledoneae/Eenzaadlobbigen_Monocotyledoneae;

rm *.bak
```