Chapter 10
Cornetto: A Combinatorial Lexical Semantic Database for Dutch

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke

10.1 Introduction

One of the goals of the STEVIN programme is the realisation of a digital infrastructure that will enforce the position of the Dutch language in the modern information and communication technology. A semantic database for Dutch is a crucial component for this infrastructure for three reasons: (1) it enables the development of semantic web applications on top of knowledge and information expressed in Dutch, (2) it provides people with access to information systems and services through their native Dutch language and (3) it will connect the Dutch language to the semantic

P. Vossen (✉) · I. Maks · H. van der Vliet
Faculty of Arts, VU University of Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands
e-mail: piek.vossen@vu.nl; e.maks@vu.nl; h.d.vander.vliet@vu.nl

R. Segers
Faculty of Sciences, VU University of Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands
e-mail: r.h.segers@vu.nl

M.-F. Moens
Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium
e-mail: Sien.Moens@cs.kuleuven.be

M. de Rijke · K. Hofmann
ISLA, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands
e-mail: derijke@uva.nl; k.hofmann@uva.nl

E. Tjong Kim Sang
Faculteit der Letteren, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK, Groningen, The Netherlands
e-mail: erik@xs4all.nl

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_10,

processing of English knowledge and information. A semantic database makes it possible to go from words to concepts and consequently, to develop technologies that access and use knowledge rather than textual representations.

At the start of STEVIN, there were two separate semantic databases for contemporary Dutch: the Referentiebestand Nederlands (RBN, [20, 21]) and the Dutch wordnet (DWN, [32, 33]). These databases contain partially overlapping and partially complementary information. More importantly, they represent different perspectives on the semantics of words: RBN follows a word-to-meaning perspective that differentiates the meanings of words in terms of their combinatoric behavior, while DWN follows a meaning-to-word perspective that defines words with the same meaning as a single concept through semantic relations between these concepts. The goal of the Cornetto project was to combine these two database into a single unique semantic resource with both the rich semantic relations taken from DWN and the typical combinatoric lexical constraints, as reflected in multiword expressions, idioms, collocations and frames taken from RBN. However, Cornetto nevertheless maintains both perspectives in the same database by representing the data as two separate but linked collections. Likewise, Cornetto can be used to view word meanings in both ways, which will eventually lead to a better and more consistent definition of the similarities and differences of the semantics and usage of words. Since the meaning-to-word view is structured according to and linked to Princeton Wordnet (PWN) [14], the semantics of the database is open to technologies developed for English. This enables transferring state-of-the-art language technologies from English to Dutch, such as semantic similarity measurement, query expansion and automatic word-sense-disambiguation.

The Cornetto database[1] was built by automatically aligning the word meanings of both databases on the basis of the overlapping information and next revising these mappings using an editor that was developed during the project. The final database contains over 92K lemmas (70K nouns, 9K verbs, 12K adjectives and 73 adverbs) corresponding to 118K word meanings. Through the alignment with PWN, ontological and domain labels were imported. In addition to the database, there is a toolkit for the acquisition of new concepts and relations, and the tuning and extraction of a domain specific sub-lexicon from a compiled corpus. Such a sub-lexicon is extracted for the domain of financial law. The Cornetto database is owned by the Dutch Language Union (Nederlandse Taalunie, NTU) and is free of charge for research.[2]

The remainder of this article is organised as follows, in Sect. 10.2 we describe work related to combining lexical resources. In Sect. 10.3 we specify the design of the database and in Sect. 10.4 we elaborate on the techniques that have been used to align RBN and DWN. In Sect. 10.5 we explain the manual editing phase

[1] http://www2.let.vu.nl/oz/cltl/cornetto/index.html

[2] Licenses can be obtained from: http://www.tst-centrale.org/nl/producten/lexica/cornetto/7-56. An external evaluation was carried out by Polderland [10]. For commercial usage, a fee must be paid to the NTU for background data that is included.

and in Sect. 10.6 we present the qualitative and quantitative results. Additionally, in Sect. 10.7 we present two acquisition toolkits that have been developed. In Sect. 10.8 we present an overview of the current use of Cornetto and finally in Sect. 10.9 we conclude with observations and lessons learned.

10.2 Related Work

In order to optimise the reusability of lexical knowledge in various resources, combining these resources becomes crucial. Many attempts involve combining lexical resources with morpho-syntactic information, e.g. [5, 12, 23, 25]. This is however a different task than matching semantic resources because it involves a finite set of specifications of limited morpho-syntactic properties instead of a large set of concepts. Once these morpho-syntactic specifications are aligned, matching lexical entries is rather trivial. One of the first approaches to semantically align lexicons was proposed in the Acquilex project [1, 6], using so-called t-links across monolingual lexical knowledge bases. These knowledge bases contain very rich and detailed typed-feature structure representations, which make matching across the specifications a form of unification. The actual links were only generated for small specialised lexicons.

One of the earliest attempts of large scale alignment was done in the EuroWordNet project [32]. In this case, wordnets in other languages are aligned with the PWN using the semantic relations in each and translations from a bilingual dictionary [13] and [2]. Although this type of alignment is cross-lingual, we used similar techniques for Cornetto but in a monolingual context and using less semantic relations. Other work on large scale alignment for monolingual resources is described in [31] and [30] for ontology alignment. This is a relatively easy task due to the rich hierarchical structure and the lack of polysemy. More complex is the work of [19], who try to align FrameNet [3] and PWN. This type of alignment comes closer to the problem addressed in Cornetto, since both are large monolingual resources with detailed descriptions of different meanings (high polysemy) and having different semantic structures.

The Cornetto project is also related to more recent work on the development of the ISO standard for lexical resources (LMF[3], ISO-24613:2008) and Wordnet-LMF [29]. Especially, Wordnet-LMF, an extension of LMF to include wordnet data, benefited from the work done in Cornetto. In Cornetto, separate collections and representations are maintained for the RBN part and the DWN part. The RBN part can be converted to an LMF representation for word meanings, while the DWN part can be structured as Wordnet-LMF, combining the benefits of both.

[3] www.lexicalmarkupframework.org/

10.3 The Design of the Database

Both DWN and RBN are semantic lexical resources. RBN uses a traditional structure of form-meaning pairs, so-called Lexical Units. Lexical Units (LUs) are word senses in the lexical semantic tradition. They contain the linguistic knowledge that is needed to properly use the word in a specific meaning in a language. Since RBN follows a word-to-meaning view, the semantic and combinatoric information for each meaning typically clarify the differences across the meanings. RBN likewise focusses on the polysemy of words and typically follows an approach to represent condensed and generalised meanings from which more specific ones can be derived.

On the other hand, DWN is organised around the notion of *synsets*. Synsets are sets of synonyms that represent a single concept as defined by [14], e.g. *box* and *luidspreker* in Dutch are synonyms for *loud speaker*. Synsets are conceptual units based the lexicalisations in a language.[4] In Wordnet, concepts are defined in a graph by lexical semantic relations, such as hypernyms (broader term), hyponyms (narrower term), role relations. Typically in Wordnet, information is provided for the synset as a whole and not for the individual synonyms, thus presenting a meaning-to-word view on a lexical database and focussing on the similarities of word meanings. For example, word meanings that are synonyms have a single gloss or definition in Wordnet but have separate definitions in RBN as different lexical units. From a Wordnet point of view, the definitions of LUs from the same synset should be semantically equivalent and the LUs of a single word should belong to different synsets. From a RBN point of view, the LUs of a single word typically differ in terms of connotation, pragmatics, syntax *and* semantics but synonymous words of the same synset can be differentiated along connotation, pragmatics and syntax but not semantics.

Outside the lexicon, an ontology provides a third layer of meaning. In Cornetto, SUMO [24] has been used as the ontological framework. SUMO provides good coverage, is publicly available, and all synsets in PWN are mapped to it. Through the equivalence relations from DWN to PWN, mappings to SUMO can be imported automatically.[5] The concepts in an ontology are referred to as Terms. Terms represent types that can be combined in a knowledge representation language to form axioms. In principle, Terms are defined independently of language but according to principles of logic. In Cornetto, the ontology represents an independent anchoring of the pure relational meaning in Wordnet. The ontology is a formal framework that can be used to constrain and validate the implicit semantic statements of the lexical semantic structures, both for LUs and synsets. Further, the semantic anchoring to the ontology contributes to the development of semantic web applications for which language-specific lexicalisations of ontological types are useful.

[4] As such, Wordnets for different languages show a certain level of idiosyncrasy.

[5] For more information about SUMO please refer to http://www.ontologyportal.org/

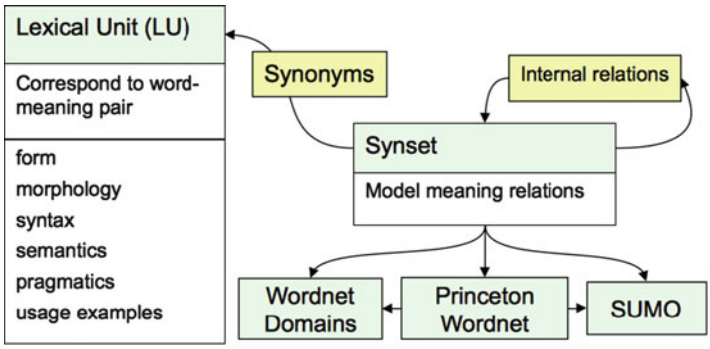

Fig. 10.1 Data collections in the Cornetto Database

A fourth layer is represented by Wordnet Domains [22]. Domains represent clusters of concepts that are related by a shared area of interest, such as *sport*, *education* or *politics*. Whereas different instruments can be subclasses of the same ontological Term (e.g. *tank* and *ambulance* are both of the type *Vehicle*), they may belong to different Domains (e.g. *military* and *medical*).

The Cornetto database (CDB) thus consists of 4 layers of information represented in two collections:

1. Collection of Lexical Units (LU), mainly derived from the RBN
2. Collection of Synsets, derived from DWN with mappings to PWN
3. Mappings to Terms and axioms in SUMO
4. Mappings to Domains in Wordnet Domains

Figure 10.1 shows an overview of the different data structures and their relations. There may be LUs that do not occur in synsets but there are no synonyms in synsets that are not LUs. The synsets are organised by means of internal relations such as hypernyms, while the LUs provide rich information on morphology, syntax and pragmatics. The synsets also point to external sources: the Princeton Wordnet (PWN), Wordnet domains (DM) and the SUMO ontology. The Cornetto database is implemented in the Dictionary Editor and Browser (DEB II) platform [18], while the raw XML files are distributed by the TST centrale. The XML Schema file for the data can be downloaded from the Cornetto website.

Figure 10.2 provides a simplified overview of the interplay between the different data structures. Here, four meanings of *band* are defined according to their semantic relations in DWN, RBN, SUMO and Wordnet Domains. Black arrows represent hypernym relations while the dashed arrows represent other semantic relations such as a Mero-Member between ‘music group’ and ‘musician’. Note that the hypernym of each synset for *band* is similar to SUMO terms, e.g. *middel* (device) and Device. However, the SUMO terms are fully axiomatised externally, while the implications of the hypernym relation remain implicit.

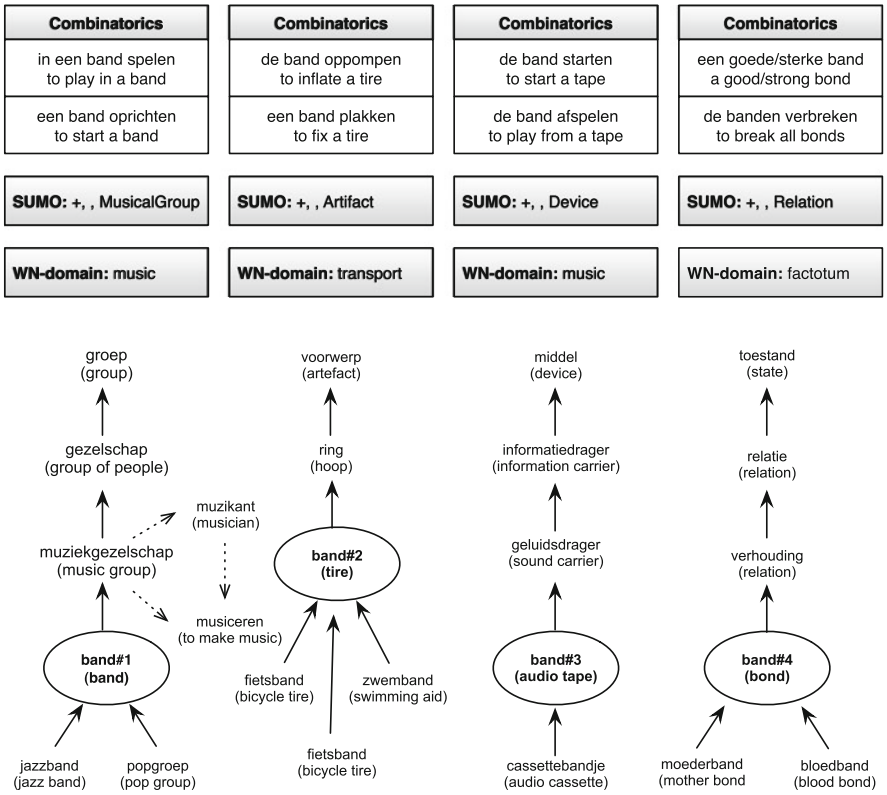

Fig. 10.2 Simplified example of the combinatorics and semantic relations for the word *band*

In the next sections, we describe the data collections for the synsets, the lexical units and the mappings to SUMO terms in more detail.

10.3.1 Lexical Units

The data structure for the LUs is implemented as a list; every LU element has an unique identifier or c_lu_id. The database for LUs contains structures to represent the form, syntactic and morphological information, semantics, pragmatics, and usage examples. An example of the XML structure for the first sense of the noun *band* (tire) is shown in Fig. 10.3. The xml of this LU contains basic morpho-syntactic information (lines 3–8), some semantics (lines 11–15) and additional examples on the combinatorial behaviour of the word such as the lexical collocation *de band oppompen* (to inflate a tire) at line 41, and an idiomatic usage: *uit de band springen* (excessive behavior) at line 20.

```
<cdb_lu c_seq_nr="1" type="swu" is_complete="true" c_lu_id="r_n-5873">
    <form form-cat="noun" form-spelling="band"/>
    <morphology_noun>
        <morpho-type>simpmorph</morpho-type>
        <morpho-plurforms><morpho-plurform>banden</morpho-plurform>
        </morpho-plurforms>
    </morphology_noun>
    <syntax_noun><sy-gender>m</sy-gender> <sy-article>de</sy-article>
    </syntax_noun>
    <semantics_noun>
        <sem-reference>common</sem-reference>
        <sem-countability>count</sem-countability>
        <sem-type>artefact</sem-type>
        <sem-subclass>vervoermiddel (deel v.)</sem-subclass>
        <sem-resume>om een wiel</sem-resume>
    </semantics_noun>
    <examples>
    <example r_ex_id="37490">
        <form_example>
            <canonicalform>uit de band springen</canonicalform>
            <category>vp</category>
        </form_example>
        <syntax_example>
            <sy-type>fixed</sy-type>
            <sy-subtype>idiom</sy-subtype>
            <sy-combi>
                <sy-combipair>
                    <sy-combiword>uit</sy-combiword> <sy-combicat>prep</sy-combicat>
                </sy-combipair>
                <sy-combipair>
                    <sy-combiword>springen</sy-combiword> <sy-combicat>verb</sy-combicat>
                </sy-combipair>
            </sy-combi>
        </syntax_example>
        <semantics_example>
            <sem-meaningdescription>zich laten gaan</sem-meaningdescription>
        </semantics_example>
    </example>
    <example r_ex_id="37491">
        <form_example>
            <canonicalform>de band oppompen</canonicalform>
            <category>vp</category>
        </form_example>
        <syntax_example>
            <sy-type>fixed</sy-type>
            <sy-subtype>lexcol</sy-subtype>
            <sy-combi>
                <sy-combipair>
                    <sy-combiword>oppompen</sy-combiword> <sy-combicat>verb</sy-combicat>
                </sy-combipair>
            </sy-combi>
            </syntax_example>
        <semantics_example>
            <sem-meaningdescription>met een pomp lucht blazen in een rubber band
                zodat hij harder wordt</sem-meaningdescription>
            <sem-lc-collocator>causeupgra</sem-lc-collocator>
        </semantics_example>
    </example>
</examples>
</cdb_lu>
```

Fig. 10.3 Shortened example of the XML structure for the lexical unit *band*

For nouns, the morpho-syntactic information is relatively simple. Figure 10.4 shows the rich information provided for verbs, illustrated by the LU *oppompen* (to inflate). The syntax field (lines 12–16) specifies the transitivity, valency and complementation of this verb. The semantics field provides information about the caseframe (lines 20–28); *oppompen* is an action verb with a selection restriction on the agent (animate agent) and no further restrictions on the theme. Finally, both a canonical (line 37) and a textual example (line 38) are given with typical fillers for the theme of this verb: ‘tube’, ‘tire’ and ‘ball’. For a further description of the structure and contents, we refer to the Cornetto deliverable [11].

```
<cdb_lu c_seq_nr="1" type="swu" is_complete="true" c_lu_id="r_v-5716">
    <form form-cat="verb" form-spelling="oppompen"/>
    <morphology_verb>
        <morpho-type>phrasal </morpho-type>
        <morpho-structure >[op]pompen</morpho-structure >
        <flex-conjugation><flex-conjugationtype>regular </flex-conjugationtype>
          </flex-conjugation>
        <flex-mode>inf </flex-mode> <flex-tense>ntense </flex-tense>
        <flex-number>nnumber</flex-number><flex-person>nperson </flex-person>
    </morphology_verb>
    <syntax_verb>
        <sy-trans>tran </sy-trans><sy-separ>sch </sy-separ>
        <sy-class>main</sy-class><sy-peraux>h</sy-peraux>
        <sy-valency>di </sy-valency> <sy-reflexiv >nrefl </sy-reflexiv >
        <sy-subject>pers </sy-subject>
        <sy-complementation><sy-comp>np </sy-comp></sy-complementation>
    </syntax_verb>
    <semantics_verb>
        <sem-type>action </sem-type>
        <sem-caseframe>
            <caseframe>action2 </caseframe>
            <args>
                <arg><caserole>agent </caserole><selrestrole>agentanimate</selrestrole>
                  <synset_list/> </arg>
                <arg><caserole>theme </caserole> <selrestrole>themenselres </selrestrole>
                  <synset_list/></arg>
            </args>
        </sem-caseframe>
        <sem-resume>vol lucht blazen </sem-resume>
    </semantics_verb>
      <pragmatics>
    <prag-domain general="true" subjectfield="tech"/>
    </pragmatics>
      <examples>
          <example r_ex_id="13374">
          <form_example>
            <canonicalform>een fietsband/tube/bal oppompen</canonicalform>
            <textualform>hij heeft zijn fietsbanden nog eens stevig opgepompt en
                                    zijn ketting goed gesmeerd</textualform>
            <category>vp</category>
            <text-category>s</text-category>
          </form_example>
          <syntax_example>
            <sy-type>free </sy-type>
              <sy-combi>
              <sy-combipair>
                <sy-combiword>fietsband </sy-combiword> <sy-combicat>noun</sy-combicat>
              </sy-combipair>
              <sy-combipair>
                <sy-combiword>tube </sy-combiword> <sy-combicat>noun</sy-combicat>
              </sy-combipair>
            </sy-combi>
            </syntax_example>
            </example>
    </examples>
</cdb_lu>
```

Fig. 10.4 Shortened example of verbal lexical unit for *oppompen* (to inflate)

10.3.2 Synsets

Synsets are identified by an unique identifier or c_synset_id, which is used to reference synsets. An additional attribute, d_synset_id, links synsets to their source concepts in DWN in order to make the lookup for the alignment process more efficient. Each synset contains one or more synonyms; each of these synonym entries consists of a pointer to a LU (c_lu_id).

Figure 10.5 illustrates the structure in more detail for the synset *band*. It has *luchtband* (tire filled with air) as a synonym (lines 2–5). Further, the example shows that *band* has several semantic relations to other concepts such as a hypernym relation to *ring* (line 20) and to various instruments that apply to *tires*, such as

```
<cdb_synset c_sy_id="d_n-38252" posSpecific="NOUN_MASCULINE" d_synset_id="d_n-38252" comment="">
  <synonyms>
    <synonym status="rbn-1-dwn-1" c_cid_id="27346" c_lu_id-previewtext="luchtband:1"
      c_lu_id="r_n-22643"/>
    <synonym status="" c_cid_id="" c_lu_id-previewtext="band:1" c_lu_id="r_n-5873"/>
  </synonyms>
  <base_concept>false</base_concept>
  <definition>met lucht gevulde band voor voertuigen;om een wiel;</definition>
  <wn_internal_relations>
    <relation factive="" reversed="false" relation_name="CO_PATIENT_INSTRUMENT"
      target-previewtext="bandelichter:1, bandafnemer:1, bandwipper:1, bandenlichter:1"
      negative="false" coordinative="false" disjunctive="false" target="d_n-21407">
        <author name="piek" score="0.0" status="YES" date="19990301" source_id="d_n-38252"/>
    </relation>
    <relation factive="" reversed="false" relation_name="CO_PATIENT_INSTRUMENT"
      target-previewtext="bandrem:2" negative="false"
      coordinative="false" disjunctive="false" target="d_n-10174">
        <author name="Piek" score="0.0" status="" date="19961217" source_id="d_n-38252"/>
    </relation>
    <relation factive="" reversed="false" relation_name="HAS_HYPERONYM"
      target-previewtext="ring:2, ringetje:1"
      negative="false" coordinative="false" disjunctive="false" target="d_n-41726">
        <author name="Paul" score="0.0" status="" date="19961206" source_id="d_n-38252"/>
    </relation>
  </wn_internal_relations>
  <wn_equivalence_relations>
    <relation target20-target20Previewtext="tire:1, tyre:2" relation_name="EQ_SYNONYM"
      target15="ENG15-03192201-n" version="pwn-1.5" target30="ENG30-04440749-n"
      target20="ENG20-04269070-n">
        <author name="Laura" score="10523.0" status="YES" date="19980903" source_id=""/>
    </relation>
  </wn_equivalence_relations>
  <wn_domains>
    <dom_relation name="roxane" status="true" term="transport"/>
  </wn_domains>
  <sumo_relations>
    <ont_relation name="dwn10_pwn15_pwn20_mapping" status="false"
      relation_name="+" arg1="" arg2="Artifact"/>
  </sumo_relations>
</cdb_synset>
```

Fig. 10.5 Example of xml structure for the synset for *band* in its first sense

bandenlichter (tire lever) at line 10, and *bandrem* (tire brake) at line 15.[6] It also shows an EQ_SYNONYM relation to the English synset for *tire* at line 27, a relation to the domain *transport* at line 34 and a subclass relation (+) to the SUMO class *Artifact* at line 38.

10.3.3 SUMO Ontology Mappings

The SUMO ontology mappings provide the conceptual anchoring of the synsets and the lexical units. The mappings to Terms in SUMO have been imported from the equivalence relations of the synsets to PrincetonWordNet (PWN). Four basic relations are used in Princeton Wordnet and Cornetto:

=	The synset is equivalent to the SUMO concept
+	The synset is subsumed by the SUMO concept
@	The synset is an instance of the SUMO concept
[	The SUMO concept is subsumed by the synset

[6]For an overview of all semantic relations used in Cornetto, we refer to Cornetto deliverable D16.

The mappings from PWN to SUMO consist of two placeholders: one for the four relations (=, +, @, []) and one for the SUMO term. In Cornetto, we extended this representation with a third placeholder to define more complex mappings from synsets to the SUMO ontology. For this, the above relations have been extended with all relations defined in SUMO (version April 2006). The relation name and two arguments represent a so-called triple.[7] The arguments of the triples follow the syntax of the relation names in SUMO: the first slot is reserved for the relation, the second slot for a variable and the third slot contains either a SUMO term or an additional variable. The variables are expressed as integers, where the integer 0 is reserved to co-index with the referent of the synset that is being defined.

For example, the following expressions are possible in the Cornetto database:

1. Equality *cirkel* (circle): (=, 0, Circle)
2. Subsumption *band* (tire): (+, 0, Artifact)
3. Related *bot* (bone) : (part, 0, Skeleton)
4. Axiomatised *theewater* (tea water): ((instance, 0, Water) (instance, 1, Making) (instance, 2, Tea) (resource, 0, 1) (result, 2,1))

Relations directly imported from Princeton Wordnet will have the structure of 1 and 2. The triples in 3 and 4 are used to specify a complex mapping relation to the SUMO ontology, in case the basic mapping relations are not sufficient. This is especially the case for so-called non-rigid concepts [16], e.g. *theewater* (water used for making tea) is not a type of water but water used for some purpose. The triples given in 4 likewise indicate that the synset refers to an instance of Water rather than a subclass and that this instance is involved in the process of making Tea as a resource.[8]

10.4 Building the Database

The semantic units of the Cornetto database, whether LUs or synsets, are based on the word meaning distinctions that are made in RBN and DWN. The database is created by aligning these units while maintaining separate collections. The smallest semantic unit is used for making the alignment, which is the LU. The overall procedure for building the database consisted of (1) an automatic alignment to create mappings for the LUs from RBN and DWN and to generate the initial Cornetto collections and (2) a manual revision of the mappings. This procedure is illustrated in Fig. 10.6 for the word *koffie* (coffee). We see that it originally had four meanings in DWN and two in RBN. The two RBN meanings match with meanings 2 and 3

[7] Note that these triples should not be mistaken with RDF triples: the Cornetto ontology triples have no URIs.

[8] For further details on the SUMO mappings in Cornetto, see the deliverable. [11]

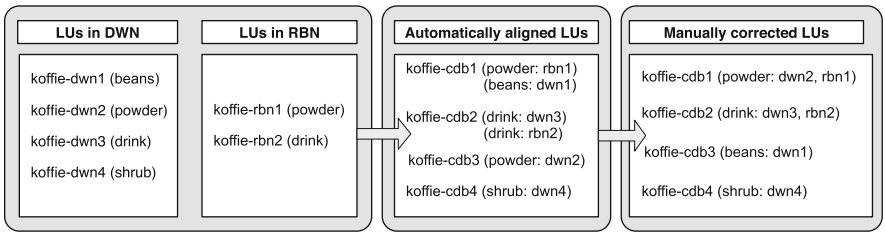

Fig. 10.6 The alignment procedure for the word *koffie* (coffee)

in DWN but the automatic procedure fails to match the second meaning (powder) of DWN and wrongly matches the first DWN meaning (beans) to the first meaning of RBN (powder). In the initial Cornetto database, we thus get four meanings for *koffie* but not all are correct. The manual revision then aligns the second DWN meaning (powder) with the first RBN meaning and creates a new LU for the first DWN meaning (beans).

The automatic alignment program initially created scored mappings across all LUs. The mappings are based on a number of heuristics taking into account: (1) the number of meanings, (2) overlapping definition words and synonyms, and (3) mappings of domains. For creating the merged database, the highest mapping relations are considered above a threshold that was empirically established from samples by eight native speakers. Precision scores varied from 54 to 97 % depending on the heuristics (more details can be found in [7]). The program created a minimal set of LUs and synsets as follows:

1. If there is a best scoring mapping between an LU in RBN and a synonym in DWN, create a single unique LU which becomes a synonym of a synset. The LU receives the ID from RBN and the synset receives the ID from DWN;
2. For all remaining mappings: do not create LUs and/or synsets in Cornetto but store additional mappings that can be accessed as weighted alternatives;
3. If there is no mapping for a LU in RBN to a synonym in DWN, create a unique LU in Cornetto with the RBN LU ID and do not create a synset for the LU in Cornetto;
4. If there is no mapping for a synonym in DWN to an LU in RBN, create (1) a synset in Cornetto with the DWN synset ID and (2) create a Cornetto LU with the DWN LU ID.

As a result, all LUs from RBN were thus copied to the Cornetto LU repository and all synsets from DWN were copied to the Cornetto synset repository. If an LU was mapped to an LU from DWN, this LU became a synonym in the DWN synset, replacing the original DWN LU. DWN LUs that could not be mapped to RBN LUs were added to the LU repository. Table 10.1 shows the degree of matching across the original resources RBN and DWN obtained through the automatic alignment. About 38 % of the LUs are matched. Almost 60 % consists of LUs from DWN not matched with RBN: mostly words not occurring in RBN. Similarly, 3,223 LUs from

Table 10.1 Number of matching and non-matching lexical units

Matches	Absolute	Relative (%)
DWN and RBN matches	35,289	37.74
LUs only in DWN	54,983	58.81
LUs only in RBN	3,223	3.45
Total	93,495	–

RBN could not be matched with a LU in DWN. For these, we did not create a new synset. The reason for this is that they often can be added manually as a synonym to an existing synset.[9]

10.5 Editing the Cornetto Database

The core Cornetto database was manually revised and checked, using an editing protocol that consisted of four main steps:

1. Manually aligning the LUs from RBN and DWN: mapping LUs to synsets, splitting, merging, deleting LUs and/or synsets
2. Adding essential information to new LUs: combinatorics, definitions, examples, etc.
3. Adding essential information to new synsets: semantic relations, SUMO mappings, Princeton Wordnet2.0 mappings
4. Manually verifying or creating mappings from existing synsets to Princeton Wordnet2.0, SUMO and WordNet Domains

The semantic information in the LUs and the synsets is complementary. As an example of the complementary combination of information, we discuss the verb *zetten* (to prepare). For the preparation of food and drinks normally the verb *maken* is used (*limonade maken* to make lemonade). This information can be found in the synset. However, in the case of making coffee or tea, one should use the verb *zetten*. The lexical constraints on phrasing the relations are not in the synsets but are provided by the LUs. Occasionally, mapping LUs and synsets raised some fundamental semantic questions. An example is the LU *brouwen* (to brew beer). This single LU corresponds to three synsets, meaning 'to brew', 'preparing a meal' and 'making plans'. The two additional meanings in DWN are metaphorical extensions; *brouwen* goes with the association of preparing, making or inventing something in an obscure way. In the synset for the concept of preparing a meal, *brouwen* is the only synonym with a clear-cut negative association. The synonyms *klaarmaken*, *toebereiden*, *bereiden* (to prepare, to make) and *koken* (to cook) are

[9]Note that the number of senses in the Cornetto database may be different from the original RBN and DWN. The RBN-based sense sequences are mostly the same, DWN-based sense sequences are mostly different.

neutral. This problem shows that the LUs and synsets differ in their perspective on word meaning. From the perspective of a LU, the aspects of meaning shared by a set of synonyms is not always an obvious meaning of a word form. As a result, aligning LUs and synsets sometimes leads to problems. In the LU of *brouwen*, the metaphorical meaning of preparing a meal was added for making the alignment with the synset possible.

In total over 10K LUs have been edited manually, corresponding to about 4,500 words that represent the most polysemous and most frequent words in the database. Another set of 509 nouns having 8 or more equivalences to PWN and 618 verbs with 5 or more equivalences were manually revised in terms of step 4. If synsets got too many mappings through the automatic mapping software, the relations are usually of low-quality and therefore also the import of the SUMO and WordNet Domain labels is unreliable.

10.6 Qualitative and Quantitative Results

In this section, an overview is presented of the main results of the alignment of DWN and RBN. First, we provide an overview of the size and coverage of the Cornetto database. Next, we discuss the quality of the alignment. Finally, we report on two task-based evaluations of the database.

Table 10.2 gives some overview statistics on the size and coverage of the database. Cornetto has about 1.6 more synsets and 1.4 as many word meanings as in the original Dutch wordnet. Our coverage compared to Princeton Wordnet is about 60 %. The average polysemy is 1.07 for nouns, 1.56 for verbs and 1.05 for adjectives.The average synset size is 1.47. The main statistics for the top-level elements of the synset data are given: the total number of synonyms, the number of internal semantic relations (like hypernyms, antonyms, meronyms, etc.), the equivalence relations to Princeton Wordnet, Wordnet Domain mappings, and SUMO mappings. Many relations are one-to-many thus exceeding the number of synsets. Furthermore, almost half of the synsets have definitions, which are derived from the resume fields of all the LUs that are synonyms of a synset.

Table 10.3 gives the main statistics for the combinatorial information related to the LUs. The 85,418 examples are subdivided into different categories: free examples illustrate the use of a LU in a wide context: the fixed examples include lexical collocations, i.e. frequent combinations with other nouns, verbs and adjectives; the grammatical collocations are frequent combinations with function words. Further, pragmatic collocations provides expressions which are associated with a fixed communicative situation.

Additional labels for the quality rates for the LU-synset mappings have been stored in a separate database. If the LU-to-synset alignment was checked manually, the quality is 100 % (10,120 alignments (9.86 %)) . These are all high frequent and high polysemous verbs, nouns and adjectives. If the mapping was not checked manually, the quality rates depend on the heuristics that underlie the mapping.

Table 10.2 Overview data Cornetto repositories

–	All	Nouns	Verbs	Adjectives
Synsets	70,370	52,845	9,017	7,689
Lexical units	119,108	85,449	17,314	15,712
Lemmas (form+POS)	92,686	70,315	9,051	12,288
Synonyms in synsets	103,762	75,475	14,138	12,914
Synonyms per synset	1.47	1.43	1.57	1.68
Senses per lemma	1.12	1.07	1.56	1.05
Definitions	35,769	25,460	6,157	4,152
Semantic relations	89,934	68,034	15,811	6,089
Equivalence relations	85,293	53,974	13,916	17,403
Domain relations	93,419	70,522	11,073	11,824
SUMO relations	70,002	46,964	12,465	10,573

Table 10.3 Overview of combinatorial information for LUs

–	All	Nouns	Verbs	Adjectives
Free examples	44,669	18,242	12,565	13,862
Lexical collocations	19,173	17,282	784	1,107
Grammatical collocations	10,407	6,869	3,160	378
Pragmatic collocations	1,472	637	565	270
Idioms	9,365	6,893	1,416	1,056
Proverbs	332	157	102	73
Total	85,418	50,080	18,592	16,746

Table 10.4 provides an overview of the quality labels that have been assigned. The number suffix after the labels indicates the reliability of the heuristic, based on a sample of 100 records per part-of-speech. For example, M-97 is a mapping with a confidence of 97 %. We see that about 53 % of the mapping records have no value for status. This means that none of the editors checked the matches and that they have not been validated in a post selection. Most of these are low frequent nouns that only occur in DWN with no match in RBN.

Additionally, two small task-based evaluations were carried out to assess the added value of the combined databases: classifying news and bootstrapping a subjectivity lexicon.

In the first task, we investigated if word combinations from Cornetto provide strong triggers for the classification of news articles using topic labels such as *sports*, *economy*, etc. [8]. The word combinations consist of lemmas combined with content words from the definitions, the examples and related synonyms and synsets. For the word *band*, we would extract combination such as *band-muziek*, *band-oppompen*, *band-moederband*. We thus extracted 60,262 records for unique forms with 396,348 word combinations. When processing news articles, the system checks every content word in the text to see if there was another content word close to it (in a window of ten words) that forms a Cornetto combination. If so, the program adds the individual content words to the index but also the combination. The assumption is that the

Table 10.4 Quality labels for *automatic* alignments

Quality label	Total	Percentage (%)	Specification
M-97	25,234	24.13	Monosemous words
B-95	4,944	4.73	Bisemous words. The first sense and second sense of RBN are aligned with the first and second sense of DWN
BM-90	4,214	4.03	Words that have one sense in RBN and two senses in DWN or vice versa. The first sense is aligned
Resume-75	1,047	1.00	Alignments based upon a substantial overlap between the RBN and DWN definitions and synonyms
D-75	2,085	1.99	Alignments of nouns that have an automatic alignment score higher than 30 %
D-58	774	0.74	Alignments of verbs that have an automatic alignment score higher than 30 %
D-55	171	0.16	Alignments of adjectives that have an automatic alignment score higher than 30 %
No-status	55,975	53.53	

combinations have a higher information value for the text than the individual words, which may be ambiguous. The baseline system was trained in a the classical way with the separate words only. As an alternative system, we also built indexes with all bigrams occurring in the training texts in addition to the individual words. For evaluation, 40 manually classified test documents were processed in the same way as the text for each of the three classification systems. The Cornetto combinations resulted in 2.8 % higher F-measure, 4.5 % higher recall and 0.5 % higher precision than the baseline. At the same time the indexes were only 1.2 times bigger by indexing word combinations using Cornetto. This means that it presents a realistic technology enhancement. The bigram system performed lower than the baseline, while its index is the biggest (4.6 times the baseline index).

The second task-based evaluation presented an algorithm that bootstraps a subjectivity lexicon from a list of initial seed examples [9]. The algorithm considers a wordnet as a graph structure where similar concepts are connected by relations such as synonymy, hyponymy, etc. We initialised the algorithm by assigning high weights to positive seed examples and low weights to negative seed examples. These weights are then propagated through the wordnet graph via the relations in the graph. After a specified number of iterations, words are ranked according to their weight. Words at the top of this ranked list are assumed to be positive and words at the bottom of the list are assumed to be negative. The algorithm was implemented and ran using two different wordnets available for Dutch: the original DWN and Cornetto. We found that using Cornetto instead of DWN resulted in a 7 % improvement of classification accuracy in the top-1500 positive words and in the top-2000 negative words. Between 70–86 % of this improvement can be attributed to the larger size of Cornetto, the remaining improvement is attributed to the larger set of relations between words.

10.7 Acquisition Toolkits

In addition to the Cornetto database, two acquisition toolkits have been developed to enhance automatic extension of the database. The first toolkit focused on automatic extraction of hypernym pairs from the web and a newspaper corpus. The second module was designed to extract domain-specific terms and collocations.

10.7.1 Acquisition Toolkit for Hypernym Pairs

For this toolkit, methods for improving the coverage of the lexical database were examined [27]. In particular, we evaluated methods for automatically extracting hypernym pairs. In texts, evidence for such a relation can be found in fixed word patterns. For example Hearst [17] explored using text patterns for finding hypernym pairs. We used the extraction approach outlined in [26] for automatically finding word pairs related by hypernymy. We extracted word pairs from the Dutch part of EuroWordNet [32] and used these as examples to train a machine learner for identifying interesting text patterns in which the pairs occurred. Next, we used the patterns that were found by the machine learner to identify other word pairs that could be related by hypernymy. We used the same machine learner as [26]: Bayesian Logistic Regression [15].

We evaluated the performance of individual text patterns and combinations of patterns on the task of extracting hypernym pairs from text. We applied the extraction method both to texts from a newspaper corpus and web text, and compared the approaches to a morphological baseline which stated that every complex noun has its final part as a hypernym, which for example predicts *bird* as hypernym of *blackbird*. We found that combined patterns outperformed individual patterns and the large web corpus outperformed the newspaper corpus. However, to our surprise none of the extraction techniques outperformed the baseline [27].

We provided the results of the newspaper texts in an online web demo [28]. The precision of the results (31 %) is in line with the state of the art, but not good enough to be useful for automatic extension of the Cornetto database. As a result , the output of the acquisition tool was not used in the construction phase of Cornetto.

10.7.2 Acquisition Toolkit for Specific Domains

In addition, we have developed a toolkit for the creation of a domain-specific lexicon containing terms and collocations [4]. For the identification of domain-specific terms and collocations, we assume large text corpora from which the terms are learned by means of statistical methods. We have experimented with common association metrics such as the likelihood ratio for a binomial distribution and a chi-square statistic, and with frequent item set mining.

The toolkit was evaluated on texts of the medical and legal domains written in Dutch. The corpora regard the medical texts of the Merck Manual, the medical encyclopedia from Spectrum and the Dutch Wikipedia articles classified in the category medicine. We created a financial law corpus obtained from the EURLex collection. As a general corpus we considered the Dutch Wikipedia pages.

For the recognition of domain-specific terms, where the association between the occurrence of a term and a domain-specific corpus was measured, the best results were obtained with a chi-square metric. Due to the lack of large and suitable training data, the results in terms of the F-1 measure hardly pass 30 %, when the extracted terms are compared with the terms found in an online medical lexicon .

Collocations can be defined as a combination of words that occur in a certain rigid order. We have extracted multi-word units, ranging from free word combinations that often co-occur to fixed idioms. The extraction is done in two steps. First, candidate collocations are identified. Second, the obtained candidates are filtered by imposing a syntactic template. In this setting, collocations could be detected such as *interne markt* (internal market), when imposing the adjectivenoun constraint, *rekening houden met* (taking into account) when imposing the noun-verb-preposition constraint, and *artikel van verordering* (article of regulation) when imposing the noun-preposition-noun constraint. According to a limited manual inspection, the obtained collocations are of good quality and the best results were obtained by frequent item set mining. Unfortunately, a complete formal evaluation by domain experts has never been performed.

10.8 Further Development of Cornetto

A number of subsequent projects have been launched that build on Cornetto:

- DutchSemCor[10] creates a sense-tagged corpus and word-sense-disambiguation software. Within this project, the Cornetto database is also extended with new, corpus-based word meanings, example sentences and semantic relations [34].
- Cornetto-LMF-RDF converts an updated version of the Cornetto database into the ISO standard LMF and the W3C standard RDF.
- Europeana[11] is a search portal on museum archives that uses the Cornetto database to provide Dutch-English cross-lingual search on meta data. Searching for *window* likewise gives results for Dutch *venster*.
- In the context of the Europeana project[12], the DWN part of Cornetto was also made available as an RDF file that consists of 792,747 triples. Cornetto-RDF is published in the linked open data cloud and linked to Wordnet W3C.

[10] www2.let.vu.nl/oz/cltl/dutchsemcor/

[11] eculture.cs.vu.nl/europeana/session/search

[12] www.europeana.eu

- FromTextToPoliticalPositions[13] develops an extension of Cornetto with fine-grained subjectivity features used for extracting political positions from text.
- SemanticsOfHistory[14] develops a domain-specific extension of Cornetto to mine historical events from text that are used in the CATCH project Agora[15].
- KYOTO[16] developed a generic fact mining platform using Cornetto as a resource. Within this project, an additional set of mappings to Princeton Wordnet was updated and edited.
- Daeso[17] is another STEVIN project that measures similarity across text. It implemented some state-of-the-art similarity measures developed for the Princeton Wordnet on top of the Cornetto database. Daeso also created a python client that can access the Cornetto database.

Since the release of Cornetto in 2008, 21 licenses have been issued by the HLT Agency in a period of 3 years.

10.9 Conclusion

The Cornetto project created a unique semantic database for Dutch and for the language community at large. The aligned information from two previously unconnected lexical databases provides a very rich database with semantic relations between concepts and traditional lexicographic information about the lexical units: e.g. combinatorics, collocations and pragmatics. By maintaining two separate collections, Cornetto provides two different views on the semantic organisation of the lexicon, which provides a form basis for studying semantics of Dutch and for developing language-technology applications. Alignment of two very differently organised lexicons proved feasible, however we argue that manual checks and editing are necessary to improve the overall quality and to solve semantic issues that stem from the different structures of the lexicons. Furthermore, the automatic acquisition toolkits provided some promising results, but also showed that acquiring a semantic lexicon from natural texts is extremely difficult for high-frequent and polysemous words and is hampered by some constraints. For instance, relations that hold between concepts are often not expressed in text as these relations are obvious for a reader.

Another major contribution is the mapping to the SUMO ontology, which allows us to differentiate rigid from non-rigid concepts and clarify the relations to entities and processes. This was taken up in subsequent projects such as KYOTO and

[13] www2.let.vu.nl/oz/cltl/t2pp/

[14] www2.let.vu.nl/oz/cltl/semhis/index.html

[15] agora.cs.vu.nl/

[16] www.kyoto-project.eu/

[17] daeso.uvt.nl/

the Global Wordnet Grid. This provides a first fundamental step towards a further formalisation of the semantics of the Dutch language and the possibility to develop semantic web applications. There is a plethora of possibilities to further extend and enrich the Cornetto database. We are considering mappings to FrameNet and creating mappings from multiword units and idioms to synsets, as well as the development of WSD systems that can assign Cornetto word meanings to words in contexts. A new version of the Cornetto database is scheduled for 2012 and includes revisions made during the DutchSemCor project [34].

References

1. Ageno A., Ribas F., Rigau G., Rodríguez H., Samiotou A.: TGE: tlinks generation environment. Proceedings of COLING'94. Kyoto, Japan (1994)
2. Atserias J., Climent S., Farreres J., Rigau G., Rodríguez H.: Combining multiple methods for the automatic construction of multilingual WordNets. Proceedings of RANLP'97. Tzigov Chark, Bulgaria (1997)
3. Baker, C., Fillmore, C., Lowe, J.: The Berkeley Framenet project. Proceedings of COLING/ACL98, Montreal, Canada (1998)
4. Boiy, E., Moens, M.-F.: Extracting domain specific collocations for the Dutch WordNet. Technical Report, Computer Science, K.U.Leuven (2008)
5. Chan, D.K., Wu, D.: Automatically merging lexicons that have incompatible part-of-speech categories. Joint SIGDAT Conference (EMNLP/VLC-99), Maryland (1999)
6. Copestake A., Briscoe E., Vossen P., Ageno A., Castelln I., Ribas F., Rigau G., Rodríguez, H., Samiotou, A.: Acquisition of lexical translation relations from MRDs. Mach Trans **9**, 3,33–3,69.
7. Cornetto deliverable D2 www2.let.vu.nl/oz/cltl/cornetto/docs/D02_Alignment of the Dutch databases in Cornetto.pdf
8. Cornetto deliverable D13 www2.let.vu.nl/oz/cltl/cornetto/docs/D03_Top-level ontology, relation constraints.pdf
9. Cornetto deliverable D14 www2.let.vu.nl/oz/cltl/cornetto/docs/D14 Tasked based evaluation subjectivity lexicon.pdf
10. Cornetto deliverable D15 www2.let.vu.nl/oz/cltl/cornetto/docs/D15_EvaluationReportCornetto.pdf
11. Cornetto deliverable D16 www2.let.vu.nl/oz/cltl/cornetto/docs/D16_Cornetto documentation (V12) v7.pdf
12. Crouch, D., King, T.H.: Unifying lexical resources. Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes. Saarbrücken, Germany (2005)
13. Farreres, X., Rigau, G., Rodríguez, H.: Using WordNet for building WordNets. Proceedings of COLING-ACL Workshop "Usage of WordNet in Natural Language Processing Systems", Montreal, Canada (1998)
14. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT, Cambridge (1998)
15. Genkin, A., Lewis, D., Madigan, D.: Large-Scale Bayesian Logistic Regression for Text Categorization. Technical report, Rutgers University, New Jersey (2004)

16. Guarino, N., Welty, C. Identity and subsumption. Green, R., Bean, C., Myaeng, S. (eds.), The Semantics of Relationships: An Interdisciplinary Perspective. Kluwer, Dordrecht, The Netherlands (2002)
17. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. Proceedings of ACL-92, Newark, DE, USA (1992)
18. Horák, A., Pala, K., Rambousek, A., Povolný, M.: DEBVisDic First Version of New Client-Server Wordnet Browsing and Editing Tool. Proceedings of GWC-06. Jeju Island, Korea (2006)
19. Laparra E., Rigau, G., Cuadros, M.: Exploring the integration of WordNet and FrameNet Proceedings of GWC2010, Mumbai, India, January 31- February 4, 2010.
20. Maks, I., Martin, W., de Meerseman, H.: RBN Manual, Vrije Universiteit, intern publication, Amsterdam (1999)
21. Martin, W.: Referentiebestand Nederlands : Documentatie Vrije Universiteit Amsterdam Internal publication (2005). Productcatalogus/RBN. www.tst.inl.nl
22. Magnini, B., Cavagliá, G.: Integrating subject field codes into WordNet. Proceedings of the LREC, Athens, Greece (2000)
23. Molinero, M.A., Sagot, B., Lionel, N.: Building a morphological and syntactic lexicon by merging various linguistic resources. Proceedings of the NODALIDA-09, Danemar (2009)
24. Niles, I., Pease, A.: Mapping WordNet to the suggested upper merged ontology. Proceedings of the IKE'03, Las Vegas, NV, USA (2003)
25. Padr, M., Bel, N., Necsulescu, S.: Towards the automatic merging of lexical resources. Proceedings of the RANLP, Hisar, Bulgaria (2011)
26. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. Proceedings of the COLING/ACL 2006, Sydney, NSW, Australia (2006)
27. Tjong Kim Sang, E., Hofmann, K.: Automatic extraction of Dutch hypernym-hyponym pairs. Proceedings of the CLIN-2006, Leuven (2007)
28. Tjong Kim Sang, E.: Cornetto Dutch Set Demo. Online web demo (2007). www.let.rug.nl/erikt/bin/setdemo.cgi
29. Soria, C., Monachini, M., Vossen, P.: Wordnet-LMf: fleshing out a standardized format for wordnet interoperability Proceedings of the IWIC2009, Stanford, CA, USA (2009)
30. Toral A., Monachini M., Soria C., Cuadros M., Rigau G., Bosma, W., Vossen, P.: Linking a domain thesaurus to WordNet and conversion to WordNet-LMF. Proceedings of the ICGL 2010, Hong Kong, China (2010)
31. Van Hage, W.: Evaluating ontology alignment techniques PhD Thesis, VU University Amsterdam (2008)
32. Vossen, P. (ed.): EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht (1998)
33. Vossen, P. (ed.): EuroWordNet General Document1, Version 3 (Final) , University of Amsterdam (2002). http://www.vossen.info/docs/2002/EWNGeneral.pdf
34. Vossen P., Gorog, A., Laan, F., Van Gompel, M., Izquierdo, R., van den Bosch, A.: DutchSemCor: building a semantically annotated corpus for Dutch. Proceedings of the eLEX2011, Bled, Slovenia, November 10–12, 2011