

The Cornetto database: architecture and alignment issues of combining lexical units, synsets and an ontology

Piek Vossen^{1,2}, Isa Maks¹, Roxane Segers¹, Hennie van der Vliet¹, Hetty van Zutphen²

¹ Faculteit der Letteren, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{[p.vossen](mailto:p.vossen@let.vu.nl), [e.maks](mailto:e.maks@let.vu.nl), [r.segers](mailto:r.segers@let.vu.nl), [h.vandervliet](mailto:h.vandervliet@let.vu.nl)}@let.vu.nl

² Irion Technologies, Delftechpark 26, 2628 XH, Delft, The Netherlands,
email: hetty.van.zutphen@irion.nl

Abstract. Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines Wordnet with Framenet-like information for Dutch. The combination of the two lexical resources (the Dutch wordnet and the Referentie Bestand Nederlands) will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. In addition to merging the Dutch lexicons, the database is also mapped to a formal ontology to provide a more solid semantic backbone. Since the database represents different traditions and perspectives of semantic organization, a key issue in the project is the alignment of concepts across the resources. This paper discusses our methodology to first automatically align the word meanings and secondly to manually revise the most critical cases.

Keywords: Wordnet, synsets, lexical units, frames, ontologies, automatic alignment

1 Introduction

Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines Wordnet with Framenet-like information for Dutch. In addition, the database is also mapped to a formal ontology to provide a more solid semantic backbone. The combination of the lexical resources will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. The database will be filled with data from the Dutch Wordnet [18] and the Referentie Bestand Nederlands [10]. The Dutch Wordnet (DWN) is similar to the Princeton Wordnet for English, and the Referentie Bestand (RBN) includes frame-like

information as in FrameNet plus other information on the combinatoric behaviour of word meanings. RBN has corpus-based examples and rich morpho-syntactic structures with complementation information. It furthermore contains many multi word expressions, both free, partly fixed and frozen expressions.

An important aspect of combining the resources is the alignment of the semantic structures. In the case of RBN these are lexical units (LUs) and in the case of DWN these are synsets. Various heuristics have been developed to do an automatic alignment. Following automatic alignment of RBN and DWN, this initial version of the Cornetto database will be further extended both automatically and manually. The resulting data structure is stored in a database that keeps separate collections for lexical units (mainly derived from RBN), for the synsets (derived from DWN) and for a formal ontology: SUMO/MILO plus extensions [15]. These 3 semantic resources represent different viewpoints and layers of linguistic, conceptual information. The alignment of the viewpoints is stored in a separate mapping table. The database is itself set up so that the formal semantic definition of meaning can be tightened for lexical units and synsets by exploiting the semantic framework of the ontology. At the same time, we want to maintain the flexibility to have a wide coverage for a complete lexicon and to encode additional linguistic information. The resulting resource will be made freely available for research in the form of an XML database.

Combining two lexical semantic databases with different organizational principles offers the possibility to study the relations between these perspectives on a large scale. However, it also makes it more difficult to align the two databases and to come to a unified view on the lexical semantic organization and the sense distinctions of the Dutch vocabulary. In this paper, we discuss the alignment issues. In section 2, we first give an overview of the structure of the database. Section 3 describes the approach and results of the automatic alignment. Section 4, discusses the manual work of checking and improving the automatic process. This work mainly involves comparing the LUs from RBN with the synset structure of DWN. Finally, in section 5, we discuss the relation between synsets and the ontology.

2 Architecture of the Database

The Cornetto database (CDB) consists of 3 main data collections:

- Collection of Lexical Units, mainly derived from the RBN
- Collection of Synsets, mainly derived from DWN
- Collection of Terms and axioms, mainly derived from SUMO and MILO

Both DWN and RBN are semantically based lexical resources. RBN uses a traditional structure of form-meaning pairs, so-called Lexical Units [3]. Lexical Units are word senses in the lexical semantic tradition. They contain all the necessary linguistic knowledge that is needed to properly use the word in a language. Word meanings that are synonyms are separate structures (records) in RBN. They have their own specification of information, including morpho-syntax and semantics. DWN is organized around the notion of Synsets. Synsets are concepts as defined by Miller and

Fellbaum [4, 12, 13] in a relational model of meaning. They are mainly conceptual units strictly related to the lexicalization pattern of a language. Concepts are defined by lexical semantic relations.¹ Typically in Wordnet, information is provided for the synset as a whole and not for the individual word meanings. For example, in Wordnet the synset has a single gloss but the different lexical units in RBN each have their own definition. From a Wordnet point of view, the definitions of lexical units that belong to the same synset should thus semantically be compatible or synonymous.

Outside the lexicon, an ontology will provide a third layer of meaning. The Terms in an ontology represent the distinct types in a formal representation of knowledge. Terms can be combined in a knowledge representation language to form expressions of axioms. In principle, meaning is defined in the ontology independently of language but according to the principles of logic. In Cornetto, the ontology represents an independent anchoring of the 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 the lexical units and the synsets. In addition, the ontology provides a mapping of a vocabulary to a formal representation that can be used to develop semantic web applications.

In addition to the 3 data collections, a separate table of so-called Cornetto Identifiers (CIDs) is provided. These identifiers contain the relations between the lexical units and the synsets in the CDB but also to the original word senses and synsets in the RBN and DWN. In Figure 1, a single CID record is shown that contains the following records:

C_form = form of the word in Cornetto
C_seq = the sequence of sense number in Cornetto
C_lu_id = the identifier of the lexical unit in Cornetto
C_syn_id = the identifier of the synset in Cornetto
R_lu_id = the identifier of the lexical unit in RBN from which it was derived
R_seq_nr = the original sequence number or sense number in RBN
D_lu_id = the identifier of the synonym in DWN
D_syn_id = the identifier of the of the synset in DWN from which it was derived
D_seq_nr = the original sequence number or sense number in DWN

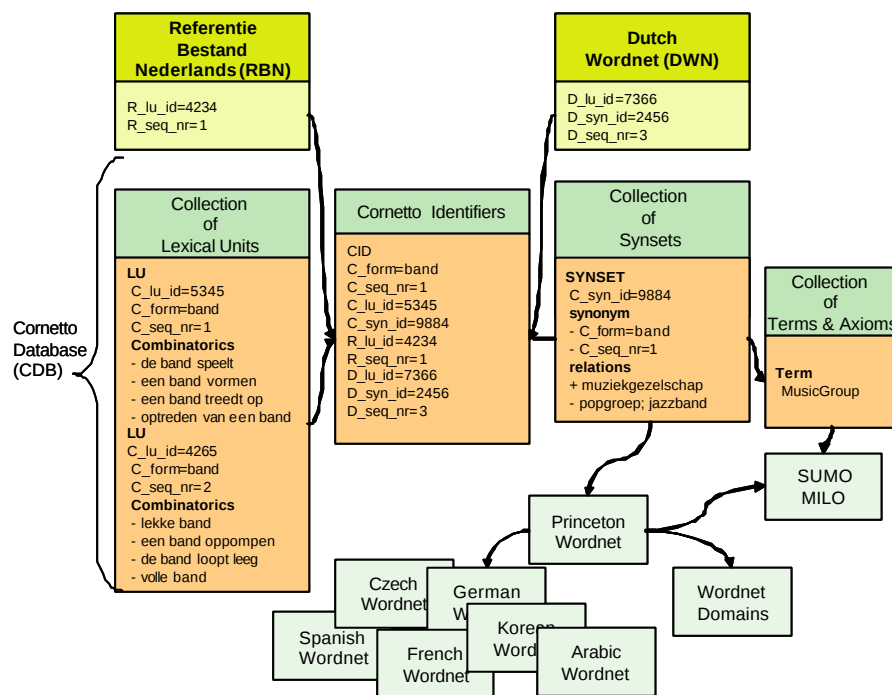
Figure 1 shows an overview of the different data structures and their relations. The different data can be divided into 3 layers of resources, from top to bottom:

- ? The RBN and DWN (at the top): the original database from which the data are derived;
- ? The Cornetto database (CDB): the ultimate database that will be built;
- ? External resources: any other resource to which the CDB will be linked, such as the Princeton Wordnet, wordnets through the Global Wordnet Association, Wordnet domains, ontologies, corpora, etc.

¹ For Cornetto, the semantic relations from EuroWordNet are taken as a starting point (Vossen1998).

The center of the CDB is formed by the table of CIDs. The CIDs tie together the separate collections of LUs and Synsets but also represent the pointers to the word meaning and synsets in the original databases: RBN and DWN and their mapping relation. As you can see in this example, the identifiers of the record match the original identifiers of synsets and lexical units in the original databases. The CIDs are just administrative records. The Cornetto data itself are stored in the collection of LUs and the collection of Synsets.

Fig. 1. Data collections in the Cornetto Database.

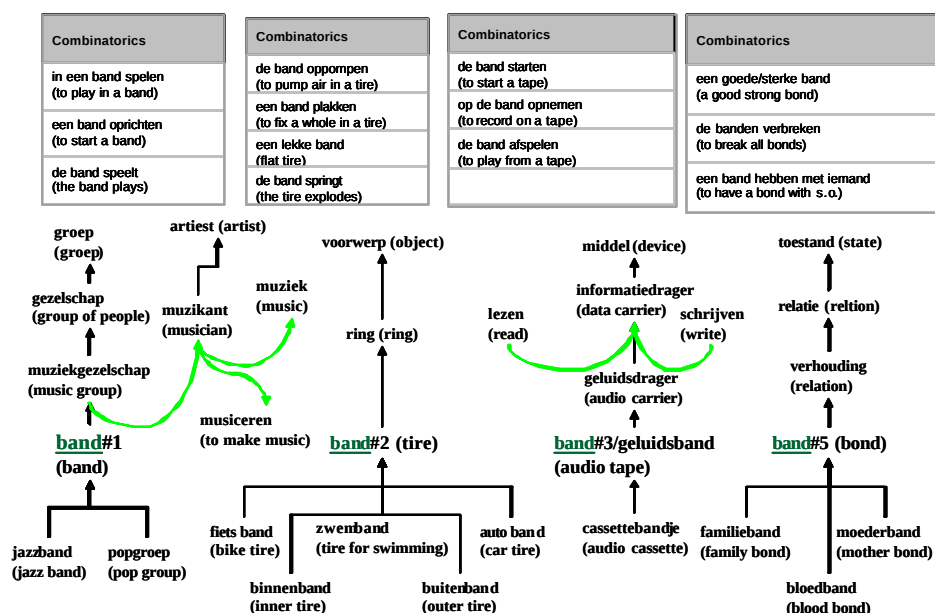


The LUs will contain semantic frame representations. The frame elements may have co-indexes with Synsets from the wordnet and/or with Terms from the ontology. This means that any semantic constraints in the frame representation can directly be related to the semantics in the other collections. Any explicit semantic relation that is expressed through a frame structure in the LU can also be represented as a conceptual semantic relation between Synsets in the Wordnet database. The Synsets in the wordnet are represented as a collection of synonyms, where each synonym is directly related to a specific LU. The conceptual relations between Synsets are backed-up by a mapping to the ontology. This can be in the form of an equivalence relation or a

subsumption relation to a Term or an expression in a knowledge representation language. Finally, a separate equivalence relation is provided to one or more synsets in the Princeton Wordnet.

The Cornetto database provides unique opportunities for innovative NLP applications. The LUs contain combinatoric information and the synsets place these words within a semantic network. Figure 2 shows an example of this combination for several meanings of the word *band*: with meanings as *musical band*, as *a tube or tire filled with air*, a *magnetic band*, and a *relationship*. The semantic network position of the word is depicted in separate wordnet fragments, relating the meanings to hypernyms, hyponyms and other related concepts. Above each fragment, we list the framelike combinatoric information that is given in RBN for these different meanings. A *musical band* is started, performs, a *tube or tire* is inflated, can leak, can blow, or you can fix it, etc. Each of these examples not only illustrates a typical conceptual usage or interaction but also the particular wording of it in Dutch. From these combinations, Dutch speakers immediately know what meaning of the word *band* applies. These typical examples can be used for the disambiguation of occurrences in text. Moreover, the same contexts can also be used for other words related to these meanings. We can easily extend the examples of *band* as a tire/tube to the hyponyms *fietsband* (bike tire) and *autoband* (car tire) and the examples of *band* as a relationship to the hypernym *verhouding* (affair) and *relatie* (relation).

Fig. 2. Combinatorics and semantics combined.



Another example, where combinatorics and semantic network relations are combined, relates to *drinks*. In Dutch, the preparation of drinks is usually referred to by the general verb *maken* (*to prepare*). However, in the case of *koffie* (*coffee*) and *thee* (*tea*), another specific verb is used: *zetten*. So, you typically use the phrases *koffie zetten* and *thee zetten* (*to make coffee or tea*) but you use the standard phrase *limonade maken* (*to make lemonade*) in Dutch. This example illustrates that conceptual combinations and constraints that are encoded in the wordnet or the ontology, do not explain the proper and most intuitive way of phrasing relations. The benefits of combining resources in this way are however only possible if the word meanings, representing concepts are properly aligned in the database. This is discussed in the next sections.

3 Aligning automatically RBN with DWN

To create the initial database, the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch part of EuroWordNet (DWN) have been automatically aligned. The word *koffie* for example has 2 word meanings in RBN (*drink and beans*) and 4 word meanings in DWN (*drink, bush, powder and beans*). This can result in 4, 5, or 6 distinct meanings in the Cornetto database depending on the degree of matching across these meanings. This alignment is different from aligning WordNet synsets because RBN is not structured in synsets. For measuring the match, we used all the semantic information that was available. Since DWN originates from the Van Dale database VLIS, we could use the definitions and domain labels from that database. The domain labels from RBN and VLIS have been aligned separately by first cleaning up the labels manually (e.g., *pol* and *politiek* can be merged) and then measuring the overlap in vocabulary associated with each domain. The overlap was expressed using a correlation figure for each domain in the matrix with each other domain. Domain labels across DWN and RBN do not require an exact match. Instead, the scores of the correlation matrix can be used for associating them. Overlap of definitions was based on the overlapping normalized content words relative to the total number of content words. For other features, such as part-of-speech, we manually defined the relations across the resources.

We only consider a possible match between words with the same orthographic form and the same part-of-speech. The strategies used to determine which word meanings can be aligned are:

1. The word has one meaning and no synonyms in both RBN and DWN
2. The word has one meaning in both RBN and DWN
3. The word has one meaning in RBN and more than one meaning in DWN
4. The word has one meaning in DWN and more in RBN
5. If the broader term (BT) of a set of words is linked, all words which are under that BT in the semantic hierarchy and which have the same form are linked
6. If some narrow term (NT) in the semantic hierarchy is related, siblings of that NT that have the same form are also linked.
7. Word meanings that have a linked domain, are linked

8. Word meanings with definitions in which one in every three content words is the same (there must be more than one match) are linked.

Each of these heuristics will result in a score for all possible mappings between word meanings. In the case of *koffie*, we thus will have 8 possible matches. The number of links found per strategy is shown in Table 1. To weigh the heuristics, we manually evaluated each heuristics. Of the results of each strategy, a sample was made of 100 records. Each sample was checked by 8 persons (6 staff and 2 students). For each record, the word form, part-of-speech and the definition was shown for both RBN and DWN (taken from VLIS). The testers had to determine whether the definitions described the same meaning of the word or not. The results of the tests were averaged, resulting in a percentage of items which were considered good links. The averages per strategy are shown in Table 1.

Table 1. Results for aligning strategies

	Conf.	Dev.	Factor	LINKS	
1: 1 RBN & 1 DWN meaning, no synonyms	97.1	4,9	3	9936	8,1%
2: 1 RBN & 1 DWN meaning	88.5	8,6	3	25366	20,8%
3: 1 RBN & >1 DWN meaning	53.9	8,1	1	22892	18,7%
4: >1 RBN & 1 DWN meaning	68.2	17,2	1	1357	1,1%
5: overlapping hyperonym word	85.3	23,3	2	7305	6,0%
6: overlapping hyponyms	74.6	22,1	2	21691	17,7%
7: overlapping domain-clusters	70.2	15,5	2	11008	9,0%
8: overlapping definition words	91.6	7,8	3	22664	18,5%

The minimal precision is 53.9 and the highest precision is 97.1. Fortunately, the low precision heuristics also have a low recall. On the basis of these results, the strategies were ranked: some were considered very good, some were considered average, and some were considered relatively poor. The ranking factors per strategy are:

- Strategies 1, 2 and 8 get factor 3
- Strategies 5, 6 and 7 get factor 2
- Strategies 3 and 4 get factor 1

A factor 3 means that it counts 3 times as strong as factor 1. It is thus considered to be a better indication of a link than factor 2 and factor 1, where factor 1 is the weakest score. The ranking factor is used to determine the score of a link. The score of the link is determined by the number of strategies that apply and the ranking factor of the strategies. In total, 136K linking records are stored in the Cornetto database. Within

the database, only the highest scoring links are used to connect WordNet meanings to synsets. There are 58K top-scoring links, representing 41K word meanings. In total 47K different RBN word meanings were linked, and 48K different VLIS/DWN word meanings. 19K word meanings from RBN were not linked, as well as 59K word meanings from VLIS/DWN. Note that we considered here the complete VLIS database instead of DWN. The original DWN database represented about 60% of the total VLIS database. VLIS synsets that are not part of DWN can still be useful for RBN, as long as they ultimately get connected to the synset hierarchy of DWN.

4 Aligning Manually RBN with DWN

The next alignment step is a manual process that consists of the editing of low-scoring and non existing links between lexical units and synsets. We identified four major groups of problematic cases and defined editing guidelines for them, which will be presented in the following sections. Many of the low-scoring links turned out to be, not unexpectedly, links between lexical units and synsets of very frequent and highly polysemous words (section 4.1 and 4.2). Many of the non-links, i.e. a link between a synset and an automatically created and therefore empty lexical unit or vice versa, turned out to be between adjective synsets and lexical units (section 4.3). The fourth group, the multiword expressions, is different from the others, since for these automatic alignment could only be performed for few cases (section 4.4).

4.1 Frequent polysemous verbs and nouns

The low-scoring links within the group of verb synsets and lexical units and within the group of noun synsets and lexical units are in great deal due to the difference regarding the underlying principles of meaning discrimination which plays an important role in the alignment of synsets and lexical units. We defined a set of 1000 most frequent verbs in Dutch as a set to manually verify. For nouns, we defined a similar set of 1800 words that are most polysemous (4 or more word meanings). The matching of nouns is relatively straight forward and the manual process consists mainly of correcting the choices or cases where different meanings are given in the two resources. In the latter case, we either create a new synset or add the word to an existing synset as a synonym or we provide the information in the lexical unit that is lacking. Mappings for verbs are more complicated as will be explained below.

Characteristic for the verbal LUs is that they contain detailed information on verbal complementation, event structure and combinatoric properties. For the verb *behandelen* (to treat), the complementation patterns are:

- ? np: *iemand behandelen* (to treat someone)
- ? np, pp: *iemand aan/ voor/ tegen/met iets behandelen*
(to treat someone for /with/ ... something)

In the representation of complementation patterns, all possible patterns are encoded. This may lead to a lot of patterns, but the result is a very explicit description of the syntactic behavior of the LU. As a rule, each pattern is worked out as an example in the combinatoric information. The corresponding event structure of *behandelen* contains the information that:

- ? this meaning of *behandelen* is an action verb.
- ? the subject np is the agent
- ? the object-np is the patient
- ? an optional pp-complement with *met (with)* is the instrument
- ? an optional pp-complement with *aan/voor/tegen (for/with/against)* is the theme

In the Dutch WordNet, these complements and roles are reflected in semantic relations:

- ? [causes] [v] genezen:2, beteren:1, herstellen:1 (*to recover*)
 - ? [involved_agent] [n] arts:1; dokter:1 (*doctor*)
 - ? [involved_patient] [n] zieke:1; patiënt:1 (*patient*)
 - ? [involved_instrument] [n] hart-longmachine:1 (*heart-long machine*)
 - ? [involved_instrument] [n] mitella:1, draagdoek:1 (*sling*)
 - ? [involved_instrument] [n] geneesmiddel:1; medicijn:1 (*medicine*)
- etc.

As long as there is a one-to-one mapping from LUs and synsets, the features of the two resources will probably match. However, difficulties arise when the mapping is not one-to-one. Frequent verbs are often very polysemous. The RBN, as the source of the LUs, tries to deal with polysemy in a systematic and efficient way. The synsets are however much more detailed on different readings. As a result, in many cases there are more synsets than LUs. In combination with the detailed information on complementation, event structure and lexical relation, this results in interesting (and time consuming!) editing problems.

A typical example of an economically created LU in combination with a detailed synset is *aflopen (to come to an end, to go off (an alarm bell), to flow down, to run down, to slope down, etc.)*. Input to the alignment are seven LUs and 13 synsets. Much of the asymmetry was caused by the fact that one of the LUs represents one basic and comprehensive meaning: *to walk to, to walk from, to walk alongside something or someone*. In DWN these are all different meanings, with different synsets. This is the result of describing lexical meaning by synsets; these three readings of *aflopen* obviously have a lot in common, but they match with different synonyms. Aligning the LUs and synsets leads to splitting the LUs and may lead to subtle changes in the complementation patterns, event structure and certainly to adapting and extending the combinatoric information. Sometimes the LUs are more detailed. In that case a synset must be split, which of course gives rise to changes in all related synsets and to new sets of lexical relations.

In every day editing of frequent verbs it is often a problem to find out the exact meaning of a verb in a synset. This is certainly the case for isolated meanings without

synonyms, forming a synset on their own, but also for frequent verbs with other frequent verb meanings in the synset. It does not help to know that *afspelen* (*to take place*) is in a synset with *passeren*, *spelen* and *geschieden* (*to happen, take place, occur*), all being ambiguous in the same way. These puzzles can often be solved by keeping a close watch on the lexical relations; especially instrument-relations are often of great help in disambiguating. However, it will be clear that alignment in the case of frequent verbs is hardly ever a matter of just confirming a suggestion for a mapping.

4.2 Nouns and semantic shifts

As is mentioned above, there are some differences in the lexicographical approach between the DWN and RBN resource for Cornetto. One important aspect is the economical distribution of LUs in the RBN, compared to the more extensive distribution of synsets. With regard to the nouns, this dissimilarity is mainly caused by the use of semantic shifts in the RBN.

A semantic shift can be defined as an aspect of a meaning that is closely connected to the central meaning. A shift can thus be seen as an extension of a meaning. Like in the RBN, the extension is not explicitly given but indicated, whereas DWN follows another approach to explicitly list these meanings. The RBN uses the semantic shift for groups of words that show the same semantic behavior. In the case of *artikel* (*article*) we find a LU with a shift that predicts that besides ‘text’, an *artikel* can also be an Artifact. This shift from Non-Dynamic to Artifact is also found consequently in LUs like *reprint* and *script*. There are about 30 different defined types of shifts that can occur in verbs, adjectives and nouns, like Process ? Action in verbs and Dynamic ? Non-dynamic in nouns. Due to the difference in approach, we expect that the matching of LUs from RBN to synonyms in DWN is more likely to be incorrect for all words labeled with a shift in RBN. We therefore decided to manually verify all the mappings for shifts. The vast majority of 4500 LUs with a semantic shift is found in nouns, on which we have decided to concentrate the manual work.

Because of the difference in approach, the DWN resource will have an extra synset for the meaning that is implied with a shift in the LU. If not, the presence of a shift might be a reason to create a new synset. This makes editing the LUs with a semantic shift a successful strategy to improve and extend the Cornetto database.

Editing an LU with a shift however, does not only mean splitting it and align it with the corresponding synset. Both resources show sometimes subtle differences in their description of a meaning, or a meaning happens to be missing in one of the resources. This means that if we want to edit the shift cases properly, we need to edit entries that contain an LU with a shift, and not just only the shift cases. This approach means that we aim at editing about 15.000 LUs and synsets, since most of the entries with a semantic shift are polysemous or will be so after editing. For these and some other edit related issues and decisions, we keep an edit log that will result in a final editing guideline.

All of this can be demonstrated by the word *bekendmaking* (*announcement*) that has one LU with a shift in RBN from Dynamic to Non-dynamic. This means that (in Dutch) an *announcement* can be a process and the result of this process. In DWN, we find a synset for each of these aspects, stating that the first one is a subclass of the

SUMO term ‘Communicating’, and the second one is equivalent to ‘Statement’. We can see this as a good argument to split the LU and define the difference in terms of the definition and the semantic relations. In almost all of the dynamic and non-dynamic cases we use the following scheme to specify the relation and differences between both synsets and LUs (fig. 3 and 4):

Fig. 3. Schemes for editing nouns with a dynamic/non-dynamic shift.

Dynamic X	
LU resume	The X-ing
LU combinatorics/example	(...)
Synset semantic relation 1	HAS_HYPERONYM ‘Y’
Synset semantic relation 2	XPOS_NEAR SYNONYM ‘X-ing’

Non-dynamic X	
LU resume	(...)
LU combinatorics/example	(...)
Synset semantic relation 1	HAS_HYPERONYM ‘X’
Synset semantic relation 2	ROLE, CAUSE, ROLE_RESULT, etc

In the case of ‘announcement’, this scheme can be filled for Dutch like this (fig. 4):

Fig. 4. An editing example for a noun with a dynamic/non-dynamic shift.

Dynamic X	announcement
LU resume	‘the announcing’
LU combinatorics/example	-
HAS_HYPERONYM	statement (<i>dynamic in Dutch</i>)
XPOS_NEAR_SYN	announcing

Non dynamic X	announcement
LU resume	‘something that has been announced’
LU combinatorics/example	-
HAS_HYPERONYM	message
ROLE_RESULT	announcing

The main advantage of editing shifts is the expansion and enrichment of the database. By creating a new LU for a synset we can add essential combinatory information and example sentences. When we add a new synset for a LU, we create new semantic relations, thus enriching the existing semantic structure of DWN. By editing clusters of the same shift type as e.g. dynamic ? non-dynamic, we can ensure consistency at the same time. Note that the label shift will be kept in both LUs: in the original LU from the RBN and in the new LU which is the explicit meaning of the shift. In this way, we can always reconstruct the original RBN approach to store a single

condensed meaning, or use the fact that there is a metonymic relation between these LUs. Furthermore, we express that there is a tight relation between these synsets.

4.3 Adjectives and fuzzy synsets

A considerable part of the adjectives is not successfully aligned by the automatic alignment procedures. This is especially due to the fact that adjective synsets have few semantic relations lacking hypemyms and hyponyms. By consequence, the automatic alignment strategies which involve broader and narrower terms, are in these cases not applicable.

Another problematic aspect of the adjective synsets is the fact that the automatically formed DWN adjective synsets are not – unlike the noun and verb synsets – edited and corrected manually. As a result, DWN adjective synsets have the following two characteristics:

- ? they are rather large and fuzzy often including words which are semantically related but not really synonymous, eg. Synset A: [dol, gek, dwaas, gaga (*mad, crazy, foolish*) achterlijk, gestoord (*retarded, disturbed*)]

The synset needs to be splitted up in at least two new synsets: A1 [dol, gek, dwaas, gaga] ‘behaving irrational’ and A2 [gestoord, achterlijk] ‘affected with insanity’.

- ? They are often quite similar to each other, e.g. Synset B [dol, dwaas, maf, (*mad, crazy, foolish*) idioot (*idiotic*), krankzinnig (*mad, insane*)...]

Although synset A includes other synonyms than synset B, they are both quite similar with respect to their meanings. They need to be partly merged into a new synset C [dol, dwaas, maf, gaga] ‘behaving irrational’ as is illustrated below (example 1).

Of course, RBN’s lexical units - with numerous corpus based examples - can be helpful in solving these problems. However, it is already mentioned that the systematic and efficient way of word sense discrimination is often not consistent with the wordnet approach. For example, the following lexical unit *kort* (*short*) shows that the RBN does not always take into consideration possible synonym or hypemym relations.

Ex. 1. RBN *kort* (*short*).

LU	Resume	Syntax	Combinatorics
Kort (short)	of time and length	attr/pred	een korte dag (a short day), een korte vakantie (a short holiday), een korte broek (short trousers) kort haar (short hair)

In this case the LU need to be split up in two LUs, distinguishing one temporal (with the combinations (1) and (2)) and one spatial sense (with the combinations (3) and (4)). Thus DWN’s semantic relations can be aligned correctly to the LUs (example 2):

Ex. 2. DWN *kort* (*short*).

Synset	Synonyms	Semantic relations
Of time	kort, kortdurend, kortstondig	Antonym : lang [1], langdurig (for a long period of time)
Of length	kort	Near-synonym: klein (small) Antonym: lang [2] (long, of relatively great length)

To be able to deal in a systematic way with these problems, we introduced the use of a semantic classification system for adjectives (Hundschnurser & Splett, Germanet). The classification regards the relation between the adjective and the modified noun. Adjectives are split up in 70 semantic classes which are organized in 15 main classes. In addition to this class, we also encode the ‘semantic orientation’ indicating a positive (+), negative (-) or neutral () connotation of the involved adjectives. Since the semantic class and the semantic orientation hold for all synonyms within the synset, it is encoded at the level of the synset.

The following example presents the aligned version – after editing both LUs and synsets – of the word *dol* (*crazy*, *fond*). We distinguished three LUs and aligned them to synsets A, B and C respectively (example 3).

Ex. 3. *dol* (LUs and Synsets).

LU	LU Resume	Syntax	Combinatorics	to Synset
1	With a strong liking for	Predicative Fixed preposition: ‘op’ (on)	Dol op kinderen (fond of children)/ dol op chocola (fond of chocolate)	A
2	Offering fun and gaiety	Attr/pred	Een dolle avond (a merry evening)	B
3	Behaving irrational	Attr/pred	Het is genoeg om dol van te worden (it is enough to drive you crazy)	C

Syn set	Synonyms	Semantic classification	Semantic orientation
A	dol, verzot, gek, verrukt	CHARACTER/BEHAVIOUR	+
B	dol, uitgelaten, jolig (<i>crazy</i> , <i>jolly</i>)	MOOD	+
C	dol, gek ,maf dwaas, gaga, geflipt (<i>crazy</i> , <i>foolish</i>)	CHARACTER/BEHAVIOUR	-

4.4 Multiword units

Special attention is paid to the encoding and alignment of multiword units. The combinatoric information in the Cornetto Database is classified into the following types: (1) free illustrative examples, (2) grammatical collocations (3) pragmatic formula (4) transparent lexical collocations (5) semi-transparent lexical collocations (6) idioms and (7) proverbs. In RBN these combinations were not included in the macrostructure, but given within the microstructure of the meaning of one particular word contained in the expression. One of the objectives of Cornetto is to introduce part of them, i.e. the fixed combinations with a reduced semantic (and often syntactic) transparency - into the macrostructure thus making it possible to align them with a synset and via the synset with the ontology. We focus on those combinations which have a reduced semantic (and often syntactic) transparency and a reduced or lack of compositionality. The following 3 types meet the criterium set for this new group:

- ? Idioms: expressions with a reduced or lack of semantic transparency (e.g. *stoken in een goed huwelijk* (drive a wedge between two people), *een rare snijboon* (an odd person)).
- ? Proverbs: completely frozen sentences.
- ? Semi-transparent lexical collocations: these are lexical collocations of which one of the combination words has got a more specific meaning or less literal meaning than its basic meaning. Therefore the whole combination has a reduced semantic transparency. (*systematische catalogus* (systematic catalogue), *open breuk* (compound fracture), *enkelvoudige breuk* (simple fracture)).

The alignment of the idioms and proverbs multiword units with the synsets will be done exclusively by hand. The alignment of the semi-transparent lexical collocations with synset hierarchy will be performed in a semi-automatic way: in most cases the synset which includes the head of the NP (*systematische catalogus*) will be the hypernym synset of the multiword unit.

With regard to their semantic description, multiword units are regarded as a sequence of words that act as a single unit. Examples (2) and (3) illustrate the encoding of a lexical collocation and an idiom respectively. The description focuses on the semantics of the whole expression: each entry consists of a canonical form, its syntactic category, a textual form (if this applies), a lexicographic definition, information regarding its use if needed, one or more examples of the construction in context. The link to the synset is realised by a pointer to a cid-entry (*c_cid_id*) and links to the individual words of the combination are realised by pointers to single word lexical units (*c_lu_id*). Morpho-syntactic information relative to the individual words is included in the description of those particular words. The pointers to the individual words are pointers to lexical units. This seems contradictory - and sometimes is - with the uncompositionality of the multiword units. However, many

multiword units are only semi-transparent and their syntactic and semantic behaviour is often related to their individual parts.

Ex. 4. Multiword unit *blinde muur* (blank wall).

Canonicalform	blinde muur (NP)
Sy-subtype	lexical collocation
meaningdescription	muur zonder ramen of deuren (a wall unbroken by windows or other openings)
C_LU_ID	muur (N) (wall)
C_LU_ID	blind (A) (blind)
Synset	[blinde muur] blank wall
Hypernym	[muur] (wall)
OntologicalType	StationaryArtifact (an artifact that has a fixed spatial location)

Ex. 5. Multiword unit *roomser dan de paus* (more Catholic than the Pope).

CanonicalForm	roomser dan de paus (AdjP)
Sy-subtype	idiom
Sem-meaningdescription	overdreven principieel (extremely principled)
Prag-Connotation	pejorative
C_LU_ID	rooms (A) (catholic)
C_LU_ID	paus (N) (pope)
Synset	roomser dan de paus
Hypernym	principieel (principled), beginselvast (consistent)
OntologicalType	TraitAttribute

5 Aligning synsets with ontology terms

A new relation is the mapping from the synset to the ontology. The ontology is seen as an independent anchoring of concepts to some formal representation that can be used for reasoning. Within the ontology, Terms are defined as disjoint Types, organized in a Type hierarchy where:

? a Type represents a class of entities that share the same essential properties.

- ? Instances of a Type belong to only a single Type: => disjoint (you cannot be both a cat & a dog)

Terms can further be combined in a knowledge representation language to form expressions of axioms (*you can be a watch dog & a bull dog*), i.e. the Knowledge Interchange Format, KIF, based on first order predicate calculus and primitive elements.

Following the OntoClean method [6, 7], identity criteria can be used to determine the set of disjunct Types. These identity criteria determine the essential properties of entities that are instances of these concepts:

- ? **Rigidity**: to what extent are properties of an entity true in all or most worlds? E.g., a *man* is always a *person* but may bear a Role like *student* only temporarily. Thus *manhood* is a rigid property while *studenthood* is anti-rigid.
- ? **Essence**: which properties of entities are essential? For example, *shape* is an essential property of *vase* but not an essential property of the clay it is made of.
- ? **Unicity**: which entities represent a whole and which entities are parts of these wholes? An *ocean* or *river* represents a whole but the *water* it contains does not.

The identity criteria are based on certain fundamental requirements. These include that the ontology is descriptive and reflects human cognition, perception, cultural imprints and social conventions (Masolo, Borgo, Gangemi, Guarino, and Oltramari 2003).

The work of Guarino and Welty (2002a, 2002b) has demonstrated that the WordNet hierarchy, when viewed as an ontology, can be improved and reduced. For example, roles such as AGENTS of processes are anti-rigid. They do not represent disjunct types in the ontology and complicate the hierarchy. As an example, consider the hyponyms of dog in WordNet, which include both types (races) like *poodle*, *Newfoundland*, and *German shepherd*, but also roles like *lapdog*, *watchdog* and *herding dog*. “Germanshepherdhood” is a rigid property, and a German shepherd will never be a Newfoundland or a poodle. But German shepherds may be herding dogs. The ontology would only list the *rigid* types of dogs (dog races): Canine => PoodleDog; NewfoundlandDog; GermanShepherdDog, etc.

The lexicon of a language then may contain words that are simply names for these types and other words that do not represent new types but represent roles (and other conceptualizations of types). For example, English *poodle*, Dutch *poedel* and Japanese *pudoru* will become simple names for the ontology type: $\Leftrightarrow$ ((instance x PoodleDog). On the other hand, English *watchdog*, the Dutch word *waakhond* and the Japanese word *banken* will be related through a KIF expression that does not involve new ontological types: $\Leftrightarrow$ ((instance x Canine) and (role x GuardingProcess)), where we assume that GuardingProcess is defined as a process in the hierarchy as well. The fact that the same expression can be used for all the three words indicates equivalence across the three languages.

In a similar way, we can use the notions of Essence and Unicity to determine which concepts are justifiably included in the type hierarchy and which ones are dependent on such types. If a language has a word to denote a lump of clay (e.g. in

Dutch *kleibrok* denotes an irregularly shaped chunk of clay), this word will not be represented by a type in the ontology because the concept it expresses does not satisfy the Essence criterion. Similarly, a word like *river water* (Dutch *rivierwater*) is not represented by a type in the ontology as it does not satisfy Unicity; such words are dependent on valid types. Satisfying the rigidity criterion, for example, is a condition for type status.

From this basic starting point, we can derive two types of mappings from synsets to the ontology [5, 19]:

- ? Synsets represent disjunct types of concepts, where they are defined as:
 - a. names of Terms;
 - b. subclasses of Terms, in case the equivalent class is not provided by the ontology
- ? Synsets represents non-rigid conceptualizations, which are defined through a KIF expression;

When we look at the different dogs in the Dutch wordnet then we see 3 types of hyponyms:

- ? bokser; corgi; loboor; mopshond; pekinees; pointer; spaniel (all dog races)
- ? pup (puppy); reu (male dog); teef (bitch)
- ? bastaard (bastard); straathond (street dog); blindengeleidehond (dog for blind people); bullebijter (nasty dog); diensthond (police dog); gashond (dog for detecting gas leaks); jachthond (hunting dog); lawinehond (aveline dog); schoothondje (lap dog); waakhond (watch dog)

The first group are names for dog races that are clearly rigid and disjunct. They represent names for Terms. The second group are words for male/female and baby dogs. They can be encoded in the same way as man, woman and child for humans. The third group refers to dogs with certain non-rigid attributes. They will thus not represent names for types but are related to the ontology by a mapping to the term Canine and the attribute that applies.

The KIF expressions are currently restricted to triplets consisting of the relation name, a first argument and a second argument. The default operator of the triplets is AND, and we assume default existential quantification of any of the variables, specified as a value of the arguments. Furthermore, we follow the convention to use a zero symbol as the variable that corresponds to the denotation of the synset being defined and any other integer for other denotations. Finally, we use the symbol $\Leftrightarrow$ for full equivalence (bidirectional subsumption). In the case of partial subsumption, we use the symbol $\Rightarrow$, meaning that the KIF expression is more general than the meaning of the synset. If no symbol is specified, we assume an exhaustive definition by the KIF expression. The symbol $\Leftrightarrow$ applies by default.

The following simplified expression can then be found in the Cornetto database for the non-rigid synset {waakhond} (watchdog): (instance, 0, Canine) (instance, 1, GuardingProcess) (role, 0, 1). This should then be read as follows:

$\Leftrightarrow$ The expression exhaustively defines the synset
 (instance, 0, Canine)
 Any referent of an expression with this synset as the head is also an
 instance of the type Canine (the special status of the zero variable),
 AND
 There exists an instance of the type Canine 0,
 AND
 (instance, 1, GuardingProcess)
 There exists an instance of the type GuardingProcess 0+1,
 AND
 (role, 0, 1)
 The entity 0 has a role relation with the entity 1.

Other expressions that we use are:

Bokser (+, 0, Canine)
 The synset {bokser} is a rigid concept which is a subclass of the type
 Canine

 hond (=, 0, Canine)
 The synset {hond} is a Dutch name for the rigid type Canine

The latter two relations are mainly imported from the SUMO mappings to the English Wordnet. In the case of {bokser} it is manually added because it is dog race that is not in the English Wordnet.

Another case of mixed hyponyms are words for *water*. In the Dutch wordnet there are over 40 words that can be used to refer to water in specific circumstances or with specific attributes. Water is in SUMO a CompoundSubstance just as other molecules. We can thus expect that the synset of *water* in Dutch matches directly to Water in SUMO, just as *zand* matches to Sand. However, *water* has 3 major meanings in the Dutch wordnet: water as liquid, water as a chemical element and a water area, while there are only two concepts in SUMO: Water as the CompoundSubstance and a WaterArea. In SUMO there is no concept for water in its liquid form, even though this is the most common concept for most people. Most of the hyponyms of *water* in the Dutch wordnet are linked to the liquid. To properly map them to the ontology, we thus first must map water as a liquid. This can be done by assigning the Attribute Liquid to the concept of Water as a CompoundSubstance:

```
(and
  (exists ?L ?W)
  (instance, ?W, Water) ,
  (instance, ?L LiquidState
  (hasAttributeinstance, ?W, ?L) )
```

In the Cornetto database, this complex KIF expression is represented by the slightly simpler relation triplets:

```
(instance, 0, Water)
(instance, 1 LiquidState)
(hasAttributeinstance, 0, 1)
```

The hyponyms of water in the Dutch wordnet can further be divided into 3 groups:

- ? Water used for a purpose: theewater (*for making tea*), koffiewater (*for making coffee*), bluswater (*for extinguishing fire*), scheerwater (*for shaving*), afwaswater (*for cleaning dishes*), waswater (*for washing*), badwater (*for bading*), koelwater (*for cooling*), spoelwater (*for flushing*), drinkwater (*for drinking*)
- ? Water occurring somewhere or originating from: putwater (*in a well*), slootwater (*in a ditch*), welwater (*out of a spring*), leidingwater, gemeentepils, kraanwater (*out of the tap*), gootwater (*in the kitchen sink or gutter*), grachtwater (*in a canal*), kwelwater (*coming from underneath a dike*), grondwater, grondwater (*in the ground*), buiswater (*on a ship*)
- ? Being the result of a process: pompwater (*being pumped away*), smeltwater, dooiwater (*melting snow and ice*), afvalwater (*waste water*), condens, condensatiewater, condenswater (*from condensation*), lekwater (*leaking water*), regenwater (*rain water*), spuiwater (*being drained for water maintenance*)

In Figure 6, you find some of the mapping expressions that are used to relate these synsets to the ontology:

Fig. 6. KIF-like mapping expressions for some hyponyms of the Dutch water.

theewater (tea water) (instance, 0, Water) (instance, 1, Human) (instance, 2, Making) (instance, 3, Tea) (agent, 1, 2) (resource, 0, 2) (result, 3, 2)	putwater (water at the bottom of well) (instance, 0, Water) (instance, 1, MineOrWell) (located, 0, 1)
--	---

bluswater (water for extinguishing fire) (instance, 0, Water) (instance, 1, Human) (instance, 2, Extinguishing) (instrument, 0, 2) (agent, 1, 2)	slootwater (in a ditch) (instance, 0, Water) (instance, 1, SmallStaticWaterArea) (part, 0, 1)
drinkwater (drinking water) (instance, 0, Water) (instance, 1, Drinking) (resource, 0, 1) (capability, 0, 1)	leidingwater, gemeentepils, kraanwater (out of the tap) (instance, 0, Water) (instance, 1, Faucet) (instance, 2, Removing) (origin, 1, 2) (patient, 0, 2)

Through the complex mappings of non-rigid synsets to the ontology, the latter can remain compact and strict. Note that the distinction between Rigid and non-Rigid does not down-grade the relevance or value of the non-rigid concepts. To the contrary, the non-rigid concepts are often more common and relevant in many situations. In the Cornetto database, we want to make the distinction between the ontology and the lexicon clearer. This means that rigid properties are defined in the ontology and non-rigid properties in the lexicon. The value of their semantics is however equal and can formally be used by combining the ontology and the lexicon.

The work on the ontology is mainly carried out manually. The mappings of the synsets to SUMO/MILO are primarily imported through the equivalence relation to the English wordnet. We used the SUMO-Wordnet mapping provided on: <http://www.ontologyportal.org/>, dated on April 2006. If there are more than one equivalence mappings with English wordnet, this may result in many to one mappings from SUMO to the synset. The mappings are manually revised traversing the Dutch wordnet hierarchy top-down so that we give priority to the most essential synsets. Furthermore, we will revise all synsets with a large number of equivalence relations or low-scoring equivalence relations. Finally, we also plan to clarify the synset-type relations for large sets of co-hyponyms as shown above for water. This work is still in progress. We do not expect this to be completed for all the synsets in this 2-year project with limited funding but we hope that a discussion on this topic can be started by working out the specification for a number of synsets and concepts.

6 Conclusion

In this paper, we presented the Cornetto project that combines three different semantic resources in a single database. Such a database presents unique opportunities to study different perspectives of meaning on a large scale and to define the relations between the different ways of defining meaning in a more strict way. We discussed the methodology of automatic and manual aligning the resources and some of the differences in encoding word-concept relations that we came across. The work on Cornetto is still ongoing and will be completed in the summer of 2008. The database and more information can be found on:

<http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm>

Acknowledgments

This research has been funded by the Netherlands Organisation for Scientific Research (NWO) via the STEVIN programme for stimulating language and speech technology in Flanders and The Netherlands.

References

1. Copestake, A., Briscoe, T.: Lexical operations in a unification-based framework. In: Pustejovsky, J. and Bergler, S. (eds.), *Lexical semantics and knowledge representation*. Proceedings of the first SIGLEX Workshop, Berkeley, pp. 101--119. Springer-Verlag, Berlin (1992).
2. Copestake, A.: Representing Lexical Polysemy. In: Klavans, J. (Ed.), *Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, pp. 21--26. Menlo Park, California (2003).
3. Cruse, D.: *Lexical semantics*. University Press, Cambridge (1986).
4. Fellbaum, C. (ed.): *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge MA (1998).
5. Fellbaum, C., Vossen, P.: Connecting the Universal to the Specific: Towards the Global Grid. In: *Proceedings of the First International Workshop on Intercultural Communication*. Reprinted in: Ishida, Toru, Fussell, Susan R. and Vossen, Piek (ed.): *Intercultural Collaboration: First International Workshop*. Lecture Notes in Computer Science, vol. 4568, pp. 1--16. Springer, New York (2007).
6. Guarino, N., Welty, C.: Identity and subsumption. In: Green, R., Bean, C., Myaeng, S. (eds.) *The Semantics of Relationships: an Interdisciplinary Perspective*. Kluwer, Dordrecht (2002).
7. Guarino, N., Welty, C.: Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2), pp. 61--65 (2002).
8. Gruber, T.R.: A translation approach to portable ontologies. In: *Knowledge Acquisition*, vol. 5 (2), pp. 199--220 (1993).

9. Horák, A., Pala, P., Rambousek, A., Povolný, M.: DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. Proceedings of the Third International Wordnet Conference (GWC-06), Jeju Island, Korea (2006).
10. Maks, I., Martin, W., Meerseman, H. de: RBN Manual, Vrije Universiteit Amsterdam (1999).
11. Magnini, B., Cavaglià, G.: Integrating subject field codes into WordNet. Proceedings of the Second International Conference Language Resources and Evaluation Conference (LREC), pp. 1413--1418. Athens (2000).
12. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.: Introduction to WordNet: An On-line lexical Database. In: International Journal of Lexicography, 3/4, pp. 235--244 (1990).
13. Miller, G. A., and Fellbaum, C. (1991). Semantic Networks of English. In: Levin, B., and Pinker, S. (eds.) Cognition, special issue, pp. 197--229 (1991).
14. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: Proceedings of FOIS 2, pp. 2-9. Ogunquit, Maine (2001).
15. Niles, I., Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the International Conference on Information and Knowledge Engineering. Las Vegas, Nevada (2003).
16. Niles, I., Terry, A.: The MILO: A general-purpose, mid-level ontology. In: Proceedings of the International Conference on Information and Knowledge Engineering. Las Vegas, Nevada (2004).
17. Pustejovsky, J.: The Generative Lexicon, MIT Press, Cambridge MA (1995).
18. Vossen, P. (ed.): EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht (1998).
19. Vossen, P., Fellbaum, C. (to appear). Universals and idiosyncrasies in multilingual wordnets. In: Boas, H. (ed.) Multilingual Lexical Resources. De Gruyter, Berlin
20. Vliet, H.D. van der: The Referentie Bestand Nederlands as a multi-purpose lexical database. International Journal of Lexicography (2007) (forthcoming).