

Connecting the Universal to the Specific: Towards the Global Grid

Christiane Fellbaum¹ and Piek Vossen²

¹ Cognitive Science Laboratory, Princeton University
fellbaum@princeton.edu*

² Irion Technologies and Faculty of Arts, Free University of Amsterdam
piek.vossen@irion.nl

Abstract. Despite their widespread use in Natural Language Processing applications, lexical databases and wordnets in particular do not yet contribute satisfactorily to the difficult problem of automatic word sense discrimination. Having built a number of lexical databases ourselves, we are keenly aware of still unresolved fundamental theoretical issues. In this paper we examine some of these questions and suggests preliminary answers concerning the nature of lexical elements and the conceptual-semantic and lexical relations that interconnect them. Our perspective is multilingual, and our goal is to formulate a proposal for a “Global Wordnet Grid” that will meet the challenge of mapping the lexicons of many languages in interesting and useful ways.

1 Introduction

Despite their widespread use in Natural Language Processing applications, lexical databases and wordnets in particular do not yet contribute satisfactorily to the difficult problem of automatic word sense discrimination. Having built a number of lexical databases ourselves, we are keenly aware of still unresolved fundamental theoretical issues. In this paper we examine some of these questions and suggests preliminary answers concerning the nature of lexical elements and the conceptual-semantic and lexical relations that interconnect them. Our perspective is multilingual, and our goal is to formulate a proposal for a “Global Wordnet Grid” that will meet the challenge of mapping the lexicons of many languages in interesting and useful ways.

The idea for a Global Wordnet Grid was born during the Third Global Wordnet Conference in Korea (January 2006). This grid will be built around a set of concepts encoded as wordnet synsets in as many languages as possible and mapped to definitions in the SUMO ontology. We envision speakers from many diverse language communities creating and contributing synsets in their language. We initially solicit encodings for the nearly 5,000 Common Base Concepts used in many current wordnet projects.³

* Work supported by the National Science Foundation and the Office of Disruptive Technology.

³ Base Concepts are expressed by synsets that occupy central positions in the wordnet structures. They tend to express general concepts relatively high up in the hierarchies

We anticipate cases of many-to-many mappings, where a given language will have more than one concept that covers the semantic space of a single Base Concept and vice versa. In other cases, a lexical item encoding a concept that is specific to a linguistic and cultural community will be included in the conceptual inventory shared by all languages, though there will be no corresponding lexemes in other languages.

Eventually, the Grid will represent the core lexicons of many languages in a form that allows further study of lexical and semantic similarities as well as disparities. Both research and applications will benefit from the Grid.⁴

2 Background: WordNet, EuroWordNet, Global WordNet

The Global Grid is a natural extension of the wordnets that have been built over the past decade. At the same time, we need to examine some fundamental assumptions that have guided past wordnets in the light of what we have learned. We begin with a brief review of the major wordnets.

2.1 WordNet

The Princeton WordNet is the first manually constructed large-scale lexical database that was widely embraced by the NLP community. WordNet was originally intended to test the feasibility of a model of human semantic memory that sought to explain economic principles of storage and retrieval of words and concepts. This model was based on the hierarchical organization of concepts expressed by nouns and the inheritance of properties (expressed by adjectives) and events (encoded by verbs) associated with these concepts. WordNet consists of four different semantic networks (one for each of the major parts of speech) that interrelated groups of cognitively synonymous words (“synsets”) via lexical and conceptual-semantic relations. For details see [14], [7], [6].

The Princeton WordNet was designed and constructed with the goal of exploring the English lexicon, without a crosslinguistic perspective. Although it was not motivated by NLP needs, the WordNet model turned out to be useful for language processing. Consequently, wordnets started to be built in other languages.

2.2 EuroWordNet

Vossen was the first expansion of WordNet into other languages [23]. Along with the construction of lexical databases for (initially) four European languages,

and to be related to many other concepts. A comparison of different wordnets led to a selection of English WordNet synsets that represent these concepts across a number of European languages. This selection is referred to as the Common Base Concepts [23].

⁴ The Grid will be publicly and freely available and no proprietary claims are made by the contributors.

the EuroWordNet design deviated from that of the Princeton WordNet and contributed several fundamental innovations that have since been adopted by dozens of additional wordnets.

To increase the connectivity among synsets, a number of new relations were defined, in particular cross-part-of-speech relations. Furthermore, all relations were marked with a feature value indicating the combinations of relations (conjunctive or disjunctive) and their directionality.

The most important difference however was the multilingual nature of the database. Each wordnet was modeled after the Princeton WordNet, having its own separate inventory of synsets and relations. In addition, the synsets of each language are linked via an “equivalence relation” to the InterLingual Index, or ILI. By means of the ILI, a synset in a given language can be mapped to a synset in any other language connected to the ILI. This design allowed the straightforward comparison of the lexicons of different languages both in terms of coverage, relations, and lexicalization patterns.

Initially, the EuroWordNet ILI was populated with the concepts (synsets) from Princeton WordNet. The reasons were mostly pragmatic — WordNet had a large coverage and was freely available. Furthermore, English was the language that was most familiar to all the European partners so judging equivalence was possible.

But several modifications and extensions of the ILI had to be considered. As WordNet was not designed as an ILI, it was often difficult to establish proper equivalence relations from the different languages. This was true even for languages that are closely related to English (like Dutch and German), and despite the fact that most European lexicons are marked by contemporary Anglo-American culture.

Compatibility between the EuroWordNet languages and the ILI with respect to lexical coverage and relations varied moreover depending on which of the two basic methods for building the European wordnets was followed:

- *Expand*: English synsets are translated into the target language and the relations are copied
- *Merge*: synsets are created for the target language, interlinked with the Princeton WordNet relations, and subsequently translated into English for mapping with ILI entries

The Expand Approach results in wordnets that are very close to the Princeton original, while the Merge Approach creates wordnets that often have a very different structure where synsets do not match straightforwardly.

2.3 Global WordNet

EWN was the first step towards the globalization of wordnets. Linguists and computer scientists in many countries started to develop WNs for their languages. Besides individual efforts, there are wordnets for entire geographic regions, such as BalkaNet [22] and the Indian WordNets (e.g., [21]). Currently, WNs exist

for some 40 languages, including dead languages like Latin and Sanskrit. For information see www.globalwordnet.org.

The authors founded the Global WordNet Association (GWA), motivated by the desire to establish and maintain community consensus concerning a common framework for the structure and design of wordnets. Another goal is to encourage the development of wordnets for all languages and to link them such that appropriate concepts are mapped across languages. The multilingual wordnets allow one to compare the lexicons of different languages on a large scale, beyond the selected few lexemes that are often considered in the investigation of particular linguistic topics. Furthermore, the availability of global wordnets opens up exciting possibilities for crosslinguistic NLP applications.

3 Challenges

The goal of mapping the lexicons of genetically and typologically unrelated languages raises the question whether there exists a universal lexicon, an inventory of concepts that are lexically encoded (or potentially encodable) by all languages. Second, what kinds of concepts does such a universal lexicon cover and how large is the common core of lexicalized concepts for most or all languages? How do language-specific lexicalizations radiate out from the core? Conversely, we ask what the differences among the lexicons of diverse languages are, whether such differences are regular and systematic, and in which areas of the lexicon they are concentrated. For the cases where individual languages show lexical gaps, we ask whether these are attributable to grammatical and structural properties or to cultural differences.

These questions inevitably lead to another, more fundamental one. What constitutes a lexeme deserving of a legitimate entry in the databases? While even linguistically naive speakers have a notion of a “word,” there is no hard definition of a word. A possible orthographic definition would state that strings of letters with an empty space on either side are words. While this would cover words such as *bank*, *sleep*, and *red*, it would wrongly leave out multiword units like *lightning rod*, *itfind out*, *word of mouth*, and *spill the beans* that constitute semantic and lexical units.⁵ Clearly, a lexical unit will merit inclusion in a database when it serves to denote an identifiable concept. But as we shall see, this criterion is less than straightforward.

Assuming at least a working definition of word, the challenge is to arrange the words of a language into a structured lexicon. Although our starting point is the WordNet model, where lexically encoded concepts are interrelated to form a semantic network, we do not take it for granted that the WordNet relations are the most suitable to represent the structure of lexicons of English or other languages. More broadly speaking, we need to ask what constitutes a valid relation among words and concepts both in a given language and crosslingually.

⁵ Note that the writing systems of many languages do not separate lexical units; clearly, this does not mean that these languages don't have words.

Finally, we explore the differences and communalities of semantic networks and ontologies. Given the notion of an ontology as a formal knowledge representation system, we ask how the lexicons of many diverse languages can be linked to an ontology such that reasoning and inferencing is enabled. Which relations should be encoded in the ontology and which ones are specific to one or more individual wordnets? Since each wordnet is also an (informal) ontology, incompatibilities between the wordnets and the formal ontology may arise. What do such mismatches tell us and what are the practical consequences for the use of wordnets for reasoning and inferencing and in Natural Language Processing?

4 What Belongs in a Universal Lexical Database?

Both formal, linguistic and informal, cultural criteria determine inclusion in the Global Grid; both turn out to be difficult to define.

Words and phrases that express available concepts must be included. *Availability* is the extent to which a word or phrase is *current* and *salient* within a language community. It affects the topics speakers talk about and the words they use to discuss these topics; it may well affect the way speakers view matters. While frequency and shared cultural background determine the degree of availability of a word or phrase, the *authority* of a speaker or a subgroup of speakers within a language community may have an effect on availability as well. For example, media have a significant influence on the words that are current; frequency counts for a given lexeme vary over time, as the newsworthiness of stories and topics grows and diminishes. Social groups determine availability and linguistic change, as studies of youth language have shown.

Each lexeme of a language is mapped onto a corresponding entry in the ontology. Languages that encode the same concept are thus mappable via the ontology. Adding the lexicons of many languages to the Global Grid will reveal which concepts are truly specific to one language only and which ones are lexicalized in other languages.

4.1 Culture-Specific Words and Concepts

In building a new wordnet and connecting to the English WordNet, one comes across cases where English has no corresponding lexicalized concept. Examples from the Dutch wordnet are the verb *klunen*, which refers to walking on skates over land to get from one frozen body of water to another. Because of different climatic, geographic, and cultural settings, this concept is specific to Dutch and not shared by many other languages (although it can be explained to, and understood by, non-Dutch speakers).

Another example is *citroenjenever*, which is a special kind of gin made with lemon skin. Unlike *klunen*, this *citroenjenever* might well be adopted by inhabitants of English-speaking countries and become a familiar concept.

Culture-specific concepts must be included in the ILI, although there may not be equivalence relations to any languages other than the one that lexicalizes such concept.

4.2 Availability and Salience

The Global Grid should include words and concepts that are available and salient in a linguistic community. This criterion may conflict with purely linguistic criteria for including words in a lexical database. Compound nouns present a case in point.

Standard lexical resources tend to follow the rule that compositional phrases like *dinner table* and *vegetable truck* need not be listed. But non-compositional compounds whose meanings is not the sum of the meanings of their components but where the entire compound is a semantic unit (*horseplay, ice luge*) must be included, as their meaning cannot be guessed even by competent speakers that are unfamiliar with these words or concepts. Non-compositionality is only one criteria for inclusion in a lexical database. Even seemingly transparent compounds like *table tennis* and *heart attack* are included in standard dictionaries (e.g., *American Heritage*), presumably because they encode frequent and salient concepts. Hence, these compounds are available to the language community, as ready-made expressions.

Compounds become established in a language community when they are frequent or salient and when their creators have a social standing that lends them what might be called “linguistic authority.” This phenomenon can be seen in the areas of science and technology, popular entertainment and commercial branding, where people introduce new terms often with the explicit intention of adding them, along with a new concept, to the lexicon. An example is Dutch *Arbeidstijdverkorting*. Although its members, *Arbeid* (“work”), *tijd* (“time”), and *verkorting* (“reduction”) suggest a straightforward compositional meaning, this compound in fact denotes a special social arrangement invented in the 1980s to create jobs, whereby people got extra spare time in exchange for a reduced salary; the reductions were intended to hire additional workers and decrease unemployment.

Conversely, the following compounds found in today’s news headlines are not to be found in any dictionary: *ministry hostages*, *celibacy ruling*, and *banana duty*. Such compounds are created on the fly, and in the context of current news stories, they are readily interpretable, yet their lifespan is limited by their newsworthiness; and only few such ad-hoc compounds will enter the lexicon on a long-term basis.

5 Lexical Mismatches as Evidence for Concepts

As in EuroWordNet, a word in any of the Grid languages will be mapped to the ILI. If the concept is also lexicalized in another Grid language, the two lexicalizations are mapped via their equivalence links to the same ILI entry. Mapping the lexicons of different languages quickly reveals cases where one language encodes a given concept and others do not. But more subtly, it shows up different ways of encoding a concept and raises the question as to what constitutes a word. We illustrate this point with a few specific cases of semantically complex verbs.

Like nouns, new verbs are regularly formed by productive processes. Different languages have different rules for conflating meaning components. Some components are free morphemes, others are bound affixed. The concepts denoted by compound verbs in one language may be expressed by simplex morphemes in other languages. While one may not want to include complex verbs in one's lexicon based on the argument that they are productive and compositional, the existence of corresponding monomorphemic lexemes in other languages argues for the conceptual status of complex verbs and hence their crosslinguistic inclusion in a multilingual resource.

5.1 Accidental Gaps

Fellbaum and Kegl examine the English verb lexicon in terms of WordNet hierarchies [8]. They argue that English has a non-lexicalized concept “eat a meal,” with its own subordinates (*dine, lunch, snack,..* and distinct from the sense of “eat” that denotes the consumption of food and has a number of manner subordinates (*nibble, munch, gulp,..*). Here, the gaps are postulated on the basis of the two semantically distinct verb groups specifying manners of eating. We assume that such gaps are language-specific and that other languages may well have distinct lexicalizations for the two superordinate *eat* concepts.

In fact, a comparison of English and Dutch verbs of cutting reveals a similar crosslinguistic asymmetry. The English verb *cut* does not specify the instrument for cutting something. Only its troponyms do: *snip, clip* imply scissors, *chop* and *hack* a large knife or an axe, etc. Dutch does not have a verb that is underspecified for the instrument, and speakers select the appropriate verb based on the default instrument, which also expresses the manner of cutting (*knippen* “cut with scissors or a scissor-like tool”, *snijden* “cut with a knife or knife-like tool”, *hakken* “to cut with an axe, or similar tool”). Thus, the lack of a Dutch superordinate verb seems accidental rather than universal.

5.2 Argument Structure Alternations

In some languages, verbal affixes change both the meaning and the argument structure of the base verb. For example, German “be-” is a locative suffix that allows the Location argument to be the direct object. Thus, verbs like *malen* (paint) and *spruehen* (spray) when prefixed with *be-* obligatorily take the entity that is being painted or sprayed (the “Location”) as their direct object:

1. Sie bemalte/bespruehte die Wand (mit Farbe)
2. She painted/sprayed/sprayed the wall (with paint)

When the material (the “Locatum”) is the direct object, the verb is in its base form:

1. Sie malte/spruehte Farbe an die Wand.
2. She painted/sprayed paint on the wall.

In English, there is no formal difference between the two meanings of such verbs, and it could be overlooked were it not for data from languages like German. However, the structure of the English WordNet forces one to reflect the differences by assuming two distinct senses that members of two different superordinates. The Location variants are manners of *cover* and the Locatum variants are manners of *apply*. A better way of representing the close semantic relation between such verb pairs would be by means of a “Perspective” relation.

6 Perspective

Both the *paint* and *spray* sentences given above can refer to one and the same event.⁶ The difference between the sentence could be referred to as one of “perspective.” To illustrate what we mean by perspective, we give another example, this one involving two lexically distinct verbs.

Converse pairs like *buy* and *sell* (that are encoded as kinds of semantic opposition in the Princeton WordNet) express the actions of different participants in the same event, a sale in this case.⁷ While the verbs and the corresponding nouns each merit their own lexical entries, we want to represent them as encodings of different perspectives on the same event. We propose to do this in the ontology.

SUMO currently distinguishes two processes as well: “Buying” and “Selling.” As in FrameNet, both events are a subclass of “FinancialTransaction” and have the same axiom that expresses a dual perspective. The SUMO-KIF representation ([15], [16]) expresses a mutual relation between two statements; one statement in which the Agent of Buying (entity x) obtains something from someone (entity y) that bears the role ORIGIN, and another statement where entity y is the Agent of the Selling process and where the entity x bears the role of DESTINATION.

The ontology thus encodes both entities as agents. A more compact encoding would be one where the two verbs *buy* and *sell* are linked to the same process and the argument structure of each verb can be co-indexed with the entities in the axiom.

Converse and reciprocal events may be encoded very differently across languages. For example, Russian has two different verbs corresponding to English *marry*, depending on whether the Agent is the bride or the groom. And whereas English encodes the difference between the activities of a teacher and a student in two different verbs, *teach* and *learn*, French uses the same verb, *apprendre*, and encodes the distinction syntactically.

⁶ It has been suggested that the Location/Locatum alternation in English is accompanied by a subtle semantic difference; Anderson states that the Location alternant implies a “holistic” reading whereby the Location is completely affected [1]. In the first sentence, this would mean that the wall is completely covered with paint. This claim has been challenged, however.

⁷ Baker et al. 1998 capture this difference by referring to two different Frame Elements — the Buyer and the Seller — of a single frame [3].

Referring to the event (sale, marriage, etc.) in the ontology allows equivalence mappings to the different languages; the encoding of distinct verbs and roles is then confined to the lexicons of each language.

7 Relations in the Global Grid

We anticipate that some lexical and semantic relations will reside in the ontology while others will be restricted to individual languages. It is an open question, subject to the investigation of a sufficiently large number of lexicons, which relations will be encoded and where. We cite a few specific cases that must be considered.

7.1 Capturing Semantic Differences via Language-Internal Relations

Some languages regularly encode semantic distinctions by means of morphology. For example, Slavic languages systematically distinguish between two members of a verb pair; one verb denotes an ongoing event and the other a completed event. English can mark perfectivity with particles, as in the phrasal verbs *eat up* and *read through*. By contrast, Romance languages tend to mark aspect by different verb conjugations on the same verb but make no distinction on the lexical level.

In Dutch, aspectual verbs can be created by prefixing a verb with *door*:

- doorademen, dooreten, doorfietsen, doorlezen, doorpraten
- (continue to breathe, eat, bike, read, talk)

An aspectual relation could be introduced for these languages that links verb synsets expressing different aspects of a given event.⁸

Another example are words marked for biological gender. While *teacher* in English is neutral and underspecified with respect to gender, many profession nouns in German, Dutch, and the Romance languages are not. In Dutch, the morphologically unmarked form *leraar* is masculine and the marked form *lerares* is feminine. While masculine and feminine nouns map to the corresponding nouns in languages that draw this distinction, both map onto a single noun in languages like English.

8 Ontology

The study of ontology goes back at least to Aristotle’s “Metaphysics,” and, as the name implies, is concerned with what exists, i.e., what concepts and categories there are in the world and what the relations among them are. Under

⁸ Note that these cases cannot be accommodated with the classical WordNet relations, such as troponymy. The aspectually marked verbs do not make encode manners of either the activity verbs (*eat*, *read*) or of aspectual verbs like *finish* or *complete*.

this definition, WordNet is an ontology, in that it records both the concepts and categories that a language encodes and the relations among them, including the hyponymy and meronymy relations proposed by Aristotle. For this reason, WordNet is often called a “lexical ontology.”⁹

Ontology has another meaning in the context of AI and Knowledge Engineering, where it is the formal statement of a logical theory. For AI systems, what “exists” is that which can be represented. A formal ontology contains definitions that associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms (see e.g., [9]). One such ontology is SUMO.

9 SUMO

SUMO, a Suggested Upper Merged Ontology [15], allows data interoperability, information search and retrieval, automated inferencing, and natural language processing. SUMO has been translated into various representation formats, but the language of development is a variant of KIF, a version of the first order predicate calculus.

SUMO consists of a set of concepts, relations, and axioms that formalize a field of interest. As an upper ontology, it is limited to concepts that are meta, generic, abstract or philosophical and hence general enough to address a wide range of domains at a high level. SUMO provides a structure upon which ontologies for specific domains such as medicine and finance can be built; the mid-level ontology MILO [17] bridges SUMO’s high-level abstractions and the low-level detail of domain ontologies.

SUMO consists of 1,000 terms and 4,000 definitional statements in first order logic language SUO-KIF (Standard Upper Ontology Knowledge Interchange Format). It is also translated into the web ontology language OWL. SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in SUO-KIF and SUMO to be expressed in many languages. SUMO has been fully mapped to the English WordNet and to wordnets in many other languages as well. Synsets map to a general SUMO term or to a term that is directly equivalent to a given synset. New formal terms are defined to cover a greater number of equivalence mappings and the definitions of the new terms depend in turn on existing fundamental concepts in SUMO. SUMO could replace the ILI created for EWN and become the ontology for all wordnets linked to the Princeton WordNet; it is currently the ontology for Arabic WordNet [4]. For example, if the Arabic word sense for *shai* (“tea”) is exhaustively defined by relations to SUMO terms, this definition can replace an equivalence relation currently encoded between the Arabic synset *shai* and the English synset *tea* in WN. If there are equivalence relations from wordnets in other languages to the

⁹ See also, for example, the Ontolinguistic research program at the University of Muenchen [20].

same PWN synset, these synsets can be linked to the same SUMO definitions, as described by Pease and Fellbaum [18].

10 The Architecture of the Grid

Using a natural language as an ILI to link the lexical and conceptual inventories of diverse languages may introduce biases and prevent the adequate representation of concepts in such languages as they must be mediated via the language of the ILI. To avoid this, the Global Wordnet Grid database will comprise a language-neutral, formal ontology as its ILI. This ontology will differ in some important respects from the ILI in EuroWordNet, which is a list of unstructured concepts derived from English WordNet:

1. The list of primitive concepts is not based on the words of a particular language but on ontological observations.
2. The concepts are related in a type-hierarchy and defined with axioms.
3. It is possible to define additional complex concepts using KIF expressions and primitive elements.

A central question that we addressed in this paper is, which concepts should be included in the ILI-ontology? We noted that the ILI-ontology must be able to encode all concepts that can be expressed in any of the Grid languages. However, the ILI need not provide a linguistic encoding — a label — for all words and expressions found in the Grid languages. We saw that many lexicalizations are transparent and systematic while others are non-compositional or seemingly ad-hoc.

We assume a reductionist view and require the ILI-ontology to contain the minimal list of concepts necessary to express equivalence across languages and to support inferencing.

Following the OntoClean method ([10], [11]), identity criteria can be used to determine what is the minimal set of concepts in all cultures where the Grid languages are used. These identity criteria determine the essential properties of entities that are instances of these concepts:

1. *rigidity*: to what extent are properties of an entity true in all worlds? E.g., a person is always a “man” but may bear a Role like “student” only temporarily; “man” is a rigid property while “student” and “father” are anti-rigid.¹⁰
2. *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.
3. *unicity*: which entities represent a whole and which entities are parts of these wholes? An “ocean” represents a whole but the “water” it contains does not.

¹⁰ See also Carlson’s discussion of individual vs. stage level predicates [5] and Pustejovsky’s discussion of Roles [19].

The identity criteria are based on certain fundamental requirements. These include that the ontology be descriptive and reflect human cognition, perception, cultural imprints and social conventions [13].

The work of Guarino and Welty has demonstrated that the WordNet hierarchy, when viewed as an ontology, can be improved and reduced [10,11]. For example, roles such as AGENTS of processes are anti-rigid. They do not represent disjunct types in the ontology and complicate the hierarchy.

Consider the hyponyms of *dog* in WordNet, which include both types (races) like *poodle*, *Newfoundland*, and *German shepherd*, but also roles like *lapdog* and *herding dog*. Germanshepherdhood is a rigid property, and a German shepherde will never be a Newfoundland or a poodle. But German shepherds may be herding dogs.

The ontology would only list the types of dogs (dog races): *Canine* \rightarrow *PoodleDog*; *NewfoundlandDog*; *GermanShepherdDog*,... If a language lexicalizes a role such as *herding dog*, the type hierarchy of the ILI is not extended, but a KIF expression is created that defines the word. An informal paraphrase of such a definition could be: (instance x Herding dog) \Leftrightarrow ((instance x Canine) and (agent x Herding)), where we assume that Herding is a process defined in the type hierarchy as well.

The type/role distinction will clear up many cases where we find mismatches or partial matches between English words and words from other languages. Earlier evaluations of mismatches in EuroWordNet [24] suggest that most mismatches can be resolved using KIF-like expressions and avoiding an extension of the type hierarchy in the ILI with new categories. Gender lexicalizations, differences in perspective, aspectual variants, etc. usually do not represent new types of concepts but can be defined with KIF expressions as well, relating them explicitly to concepts that are types.

When words in the Grid languages suggest new types, the ontological criteria can be used to decide on extensions of the type hierarchy. This is the case not only for culture-specific concepts but also for other kinds of lexicalization differences. For example, the specific ways of cutting lexicalized in Dutch are actually distinct types of processes. In this case, Dutch would be the source for the extension of event types, as the English lexicalization remain too abstract.

In summary, the proposed ILI-ontology has the following characteristics:

1. It is *minimal* so that Terms are distinguished by essential properties only (reductionist)
2. It is *comprehensive* and includes all distinct concept types of all Grid languages
3. It allows the definition of all lexicalizations that express non-essential properties of the types using KIF expressions
4. It is *logically valid* and usable for inferencing

In EuroWordNet, equivalence relations currently vary considerably. Some wordnets only have “exact” equivalence, while others also allow “near equivalence” and have many-to-many relations among synsets and the corresponding concepts in the ILI.

The ILLI-ontology we propose here will be more explicit about the meaning of the equivalence relation. Because the ontology is minimal, it will be easier to establish precise and direct equivalences from Grid languages to the ontology and likewise equivalence across languages. The multilingual Grid database will thus consist of wordnets with synsets that are either simple names for ontology types in the type hierarchy or words that relate to these types in a complex way, made explicit in a KIF expression. Note that if two Grid language wordnets create the same KIF expression, they state equivalence without an extended type hierarchy.

11 Towards the Realization of the Global Grid

We propose to take the SUMO ontology as a starting point for three reasons:

- It is consistent with many ontologies and ontological practice.
- It has been fully mapped onto WordNet.
- Like WordNet, it is freely and publicly available.

SUMO, an upper ontology, is by far not rich or large enough to replace the Princeton WordNet as an ILLI-ontology. The current mapping of SUMO to WordNet will be taken as a starting point; most of these mappings are subsumption relations to general SUMO types. The first step is therefore to extend the SUMO type hierarchy so that it becomes as rich as WordNet with respect to disjoint types. Note that not all synsets from WordNet are necessary. In fact, all WordNet synsets must be reviewed with respect to the OntoClean methodology [11] so that only rigid (and semi-rigid) concepts are preserved. All remaining synsets must be defined using KIF expressions as described earlier. For example, the English word *watchdog* would get a simple KIF expression like: $\Leftrightarrow ((\text{instance } x \text{ Canine}) \text{ and } (\text{role } x \text{ GuardingProcess}))$, where x co-indexes with the referent of the noun.

Subsequently, other languages that have already established equivalence relations with WordNet can replace these with the improved mappings to SUMO, which can be copied from the Princeton WordNet. For example, Dutch *poedel* and Japanese *pudoru* will become simple names for the type $\Leftrightarrow ((\text{instance } x \text{ Poodle}))$, because they are equivalents of WordNet synset *poodle*. Similarly, Dutch *waakhond* and Japanese *banken* would be linked to the same KIF expression as both are equivalent to *watchdog* in WordNet; the KIF expression can simply be copied.

In other cases, the equivalence relations to WordNet may require some revision as it is now possible to express certain subtle distinctions between concepts expressed in the Grid languages and corresponding ones in English that could not be expressed in the EuroWordNet model. For example, the Dutch verb *bankdrukken* will be related to the English noun *bench press*, meaning *a weightlifting exercise*, because there is no corresponding verb synset in WordNet. The part-of-speech mismatch does not allow a direct match between the Dutch and English synsets. The ontology will include a process “BenchPress”

that is not marked for part of speech; both the English noun and the Dutch verb can be linked to this same process. This does not prevent us from indicating further differences, such as aspectual meaning.

Importantly, this design makes it unnecessary to write separate KIF expressions for ontological concepts in each language — most expressions can be linked via their relations to English synsets and revisions are required in some cases only.

However, the situation will arise where synsets in Grid languages cannot be mapped to WordNet. In those cases, the concepts represented by these synsets need to be checked for adherence to OntoClean. This step may result in extensions to the type hierarchy in some cases; in other cases, the wordnet builders need to write a KIF expression clarifying the particular concept's relation to the ontology. For example, The Dutch noun *straathond* (street dog), which is not mapped to WordNet, can be defined relatively easily: \sqsubset ((instance x Canine) and (habitat x Street)), following the model of similar expressions such as *watchdog* and *herding dog*.

Synsets that are not disjunct types usually have a relatively straightforward semantic structure and the KIF expressions in many cases can be copied from similar synsets that can be identified in the ontology by browsing through the hierarchy of roles and processes.

We are aware that highly specific concepts, restricted to a given culture, may be present difficulties for providing an exhaustive and satisfactory KIF expression. A solution is to provide a definition, or “gloss,” in the Grid language, a corresponding English gloss, and the most specific superclass in the type hierarchy of the ontology.

New types could be created and built in a Wiki environment. Initially, full definitions are not necessary; it is more important to have a comprehensive list of type candidates that become more precisely defined by the community in the course of the Grid construction. Furthermore, the possibility should be explored to allow the creation of KIF expressions via a simple interface or questionnaire that makes such expressions accessible to linguists and speakers of a language unfamiliar with ontology.

12 Conclusion

The Global Wordnet Grid can only be realized in a collaborative framework among builders of wordnets from many diverse linguistic and cultural backgrounds. Its development will undoubtedly include several steps and many rounds of refinement. Throughout the development of the Global Wordnet Grid, we expect discussion and the need for revisions as more languages join and the coverage for each language increases. Mapping the lexicons of many diverse languages, and the cultural notions they encode, is bound to be a long and painful process, but also a worthwhile one. The result will be a unique database that allows for a better understanding among people from different linguistic and cultural backgrounds and opens up new possibilities for research and applications.

We think it is important that such a database is built on a large scale and that it is based on a diverse set of languages and cultures. The languages will form the empirical evidence and basis for the construction of a truly universal index.

References

1. Anderson, S.R.: On the Role of Deep Structure in Semantic Interpretation. *Foundations of Language* 7 (1982), 387-396.
2. Apresyan, J.D.: Regular Polysemy. *Linguistics* 142 (1973), 5-32.
3. Baker, C., Fillmore, C. J., and Lowe, J. B.: The Berkeley FrameNet project. *Proceedings of the COLING-ACL* (1998), Montreal, Canada.
4. Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Bertran, M., and Fellbaum, C.: The Arabic WordNet Project. *Proceedings of the Conference on Lexical Resources in the European Community* (2006), Genoa, Italy.
5. Carlson, G.: *Reference to Kinds in English*. New York: Garland Press (1980).
6. Fellbaum, C.: The English Verb Lexicon as a Semantic Net. *International Journal of Lexicography*, 3, 1990, 278-301.
7. Fellbaum, C. (Ed.): *WordNet: An Electronic Lexical Database*. MIT Press, (1998), Cambridge, MA.
8. Fellbaum, C., and Kegl, J.: Taxonomic Structure and Object Deletion in the English Verbal System. In: deJong, K., and No, Y. (Eds.), *Proceedings of the Sixth Eastern States Conference on Linguistics*, (1989), 94-103. Ohio State University, Columbus, OH.
9. Gruber, T.R.: A translation approach to portable ontologies. *Knowledge Acquisition*, 5, (1993), 199-220.
10. Guarino, N. and Welty, C.: Identity and subsumption. In: R. Green, C. Bean and S. Myaeng (Eds.), *The Semantics of Relationships: an Interdisciplinary Perspective* Kluwer (2002).
11. Guarino, N. and Welty, C.: Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2) (2002) 61-65.
12. Levin, B. 1993: *English Verb Classes and Alternations*. University of Chicago Press (1993), Chicago, IL.
13. Masolo, C., Borgo, S., Gangemi, A., Guarino, N. and Oltramari, A.: WonderWeb Deliverable D18 Ontology Library. Laboratory For Applied Ontology - IST-CNR, Trento (2003).
14. Miller, G. A. (Ed.): WordNet. Special Issue of the *International Journal of Lexicography*, 3, (1990).
15. Niles, I., and Pease, A.: Towards a Standard Upper Ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine, (2001).
16. Niles, I. and 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).
17. Niles, I. and Terry, A.: The MILO: A general-purpose, mid-level ontology. *Proceedings of the International Conference on Information and Knowledge Engineering*, (2004) Las Vegas, Nevada.
18. Pease, A., and Fellbaum, C.: Formal Ontology as Interlingua. In: Huang, C. R. and Prevot, L. (Eds.) *Ontologies and Lexical Resources*, (2007), Cambridge: Cambridge University Press.

19. Pustejovsky, J.: *The Generative Lexicon*. Cambridge, MA: MIT Press (1995).
20. Schalley, A. C. and Zaefferer, D. (Eds.): *Ontolinguistics*. Berlin: Mouton de Gruyter (2007).
21. Sinha, M., Reddy, M., and Bhattacharyya, P.: An Approach towards Construction and Application of Multilingual Indo-WordNet. *Proceedings of the Third Global Wordnet Conference*, Jeju Island, Korea, (2006).
22. Tufis, D. (Ed.): The BalkaNet Project. Special Issue of *The Romanian Journal of Information Science and Technology*, 7, (2004), 1-248.
23. Vossen, P. (Ed.): *EuroWordNet*. Kluwer, Dordrecht (1998).
24. Vossen, P., Peters, W. and Gonzalo, J.: Towards a Universal Index of Meaning. In: *Proceedings of ACL-99 Workshop, Siglex-99, Standardizing Lexical Resources*, June 21-22, University of Maryland, College Park, Maryland.p 81- 90, (1999).