Springer

Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].
If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: http://www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

ArticleTitle	Challenges for a multilingual wordnet	
Article Sub-Title		
Article CopyRight	Springer Science+Business Media B.V. (This will be the copyright line in the final PDF)	
Journal Name	Language Resources and Evaluation	
Corresponding Author	Family Name	**Fellbaum**
	Particle	
	Given Name	**Christiane**
	Suffix	
	Division	Department of Computer Science
	Organization	Princeton University
	Address	Princeton, NJ, 08540, USA
	Email	fellbaum@princeton.edu
Author	Family Name	**Vossen**
	Particle	
	Given Name	**Piek**
	Suffix	
	Division	Faculty of Arts
	Organization	VU University of Amsterdam
	Address	Amsterdam, 1081 HV, The Netherlands
	Email	
Schedule	Received	
	Revised	
	Accepted	
Abstract	Wordnets have been created in many languages, revealing both their lexical commonalities and diversity. The next challenge is to make multilingual wordnets fully interoperable. The EuroWordNet experience revealed the shortcomings of an interlingua based on a natural language. Instead, we propose a model based on the division of the lexicon and a language-independent, formal ontology that serves as the hub interlinking the language-specific lexicons. The ontology avoids the idiosyncracies of the lexicon and furthermore allows formal reasoning about the concepts it contains. We address the division of labor between ontology and lexicon. Finally, we illustrate our model in the context of a domain-specific multilingual information system based on a central ontology and interconnected wordnets in seven languages.	
Keywords (separated by '-')	Multilingual wordnets - Formal ontology - Information system	
Footnote Information		

Journal: 10579
Article: 9186

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Please confirm the section headings are correctly identified.	
2.	Please check and confirm the volume number and issue number is correctly identified for the reference Gangemi et al. (2003).	
3.	Please check and confirm the volume number, issue number and page range are correctly identified for the references Miller (1990), Tufis (2004).	
4.	Please check and confirm the page range is correctly identified for the reference Miller (1995).	

Journal : 10579 | Dispatch : 30-4-2012 | Pages : 14
Article No. : 9186 | □ LE | □ TYPESET
MS Code : LRE881 | ☑ CP | ☑ DISK

Lang Resources & Evaluation
DOI 10.1007/s10579-012-9186-z

ORIGINAL PAPER

Challenges for a multilingual wordnet

Christiane Fellbaum · Piek Vossen

Abstract Wordnets have been created in many languages, revealing both their lexical commonalities and diversity. The next challenge is to make multilingual wordnets fully interoperable. The EuroWordNet experience revealed the shortcomings of an interlingua based on a natural language. Instead, we propose a model based on the division of the lexicon and a language-independent, formal ontology that serves as the hub interlinking the language-specific lexicons. The ontology avoids the idiosyncracies of the lexicon and furthermore allows formal reasoning about the concepts it contains. We address the division of labor between ontology and lexicon. Finally, we illustrate our model in the context of a domain-specific multilingual information system based on a central ontology and interconnected wordnets in seven languages.

Keywords Multilingual wordnets · Formal ontology · Information system

1 Introduction

Digital lexical resources can store lexicons of potentially unlimited size in ways that enable flexible representations and searches. Mapping the lexical inventory of a language into a semantic network has proved to be useful for many natural language processing applications, and WordNet-style lexicography has been applied to build

A1 C. Fellbaum (✉)
A2 Department of Computer Science, Princeton University, Princeton, NJ 08540, USA
A3 e-mail: fellbaum@princeton.edu

A4 P. Vossen
A5 Faculty of Arts, VU University of Amsterdam, 1081 HV Amsterdam, The Netherlands

resources in many languages.[1] The challenge we face now is to interconnect them so as to create one multilingual database. To reflect intra- and inter-lingual meaning, we argue for the connection of lexical databases to a shared ontology, which requires us to resolve some fundamental linguistic and ontological questions. We address these in the context of an ongoing project that represents a first step in the creation of a global wordnet system.

2 The scope of a lexicon

Lexical databases do not need to obey constraints on their size, and no well-defined guidelines exist for what is to be included in the lexicon. Lexicons are idiosyncratic; they do not systematically label concepts, and the idiosyncracies are often revealed in crosslinguistic differences. Lexicons are moreover redundant, often assigning several word forms (synonyms) to a single concept. Finally, lexicons are open-ended, often extended into terminology and domain-specific vocabulary.

Because inclusion in the lexicon cannot be determined by well-defined rules, its boundaries are fuzzy. Moreover, the lexical status of many phrases and chunks is uncertain, raising the question as to what constitutes a lexeme deserving of a legitimate entry in the databases.

While even linguistically naive speakers have an intuitive notion of "word," there exists no hard definition. 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 *road, eat*, and *heavy*, it would wrongly leave out multiword units like *lightning rod, find out, word of mouth*, and *spill the beans* that constitute semantic and lexical units.[2] A first rule of thumb might state that a lexical unit will merit inclusion in a database when it serves to denote an identifiable concept. But this criterion is less than straightforward, especially when applied to multi-word units.

2.1 Compositionality, currency, salience, linguistic authority

For NLP applications, including multi-word phrases and sentence chunks besides single words may present clear processing advantages. However, even so-called "fixed expressions" are subject to lexical variation and internal modification (e.g., Fellbaum 2007). The lexical status of multiword units is usually determined on the basis of the compositionality criterion. If the meaning of the whole is the sum of the meaning of its parts, there is no reason to consider the multiword string a separate entity. Thus, fully compositional phrases like *coniferous forest* and *subtropical coniferous forest* should probably not be considered as separate fixed lexical items. By contrast, a non-compositional term like *ecological footprint* cannot be readily analyzed by speakers and must be listed in the lexicon. Additional factors, not based

[1] We will reserve the legally registered name "WordNet" for the Princeton WordNet and use "wordnet" as a generic term to denote semantic networks inspired by the Princeton WordNet.

[2] In languages whose writing systems do not separate lexical units, the notion of word is of course divorced from that of a graphemic unit.

on linguistic economy, might override the compositionality criterion. *Currency, salience,* and *speaker authority* are three such interacting factors.

Currency is the extent to which a speaker community avails itself of a word or phrase that becomes (often temporarily) salient through frequent use. While frequency and shared cultural background determine the currency of a word or phrase, the authority of a speaker or a subgroup of speakers within a language community may have an effect on a word's use as well. Thus, popular media exercise a significant influence on the words that are circulating within a speaker community; frequency counts for a given lexeme vary over time, as the newsworthiness of stories and topics grows and diminishes. Social groups determine acceptance and linguistic change, as studies of youth language have shown.

Within a specific domain, a multiword term may be particular salient, as reflected in its frequency or its function as a topic of numerous documents. Over time, such compounds may acquire the status of fixed, ready-made expressions and become a part of the lexicon of the language community. Compound terms become established in a language community when their creators and initial users have a social standing that bestows on them a kind of "linguistic authority." This phenomenon can be seen in the areas of science and technology (*mouse potato*), popular entertainment and commercial branding (*e-inkReader*), where people introduce new terms often with the wish of adding them, along with a new concept, to the lexicon.

An example of currency, salience and perhaps linguistic authority is the Dutch noun compound *arbeidstijdverkorting*. Although its members, *arbeid* ("work"), *tijd* ("time"), and and *verkorting* ("reduction") suggest a straightforward compositional meaning, this compound in fact denotes more than the mere sum of its members: a specific social arrangement dating to the 1980s intended to decrease unemployment; work hours and wages were reduced so that additional workers could be hired and new jobs could be created.

3 WordNet, EuroWordNet, global wordnet

Digital lexicography resulted in the abandonment of orthography as an organizing principle for dictionaries. Representing the lexicon as a semantic network has proved particularly useful for NLP applications, and WordNet-style resources were built for many languages. We briefly review the principles of wordnet lexicography and the development of multilingual wordnets.

The Princeton WordNet (Miller 1990, 1995; Fellbaum 1998) is a manually constructed large-scale lexical database for English. WordNet's original motivation was to test the feasibility of a model of human semantic memory that sought to explain principles of storage and retrieval of words and concepts. This model proposed a largely hierarchical organization of concepts expressed by nouns, events (encoded by verbs) and properties (expressed by adjectives). The WordNet experiment tried to determine whether the bulk of the lexicon of a language could indeed be represented in a semantic network by means of a handful of relations, inspired by the model of human memory.

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

An important semantic relation is that between words sharing the same denotational meaning, synonymy. WordNet groups synonymous words into “synsets,” the building blocks, or nodes, of the semantic network. Synsets are interlinked by means of semantic relations, such as hyponymy (the super-subordinate relation that holds between words like *building* and *garage*), meronymy (the part-whole relation that hold between words like *toe* and *foot*), antonymy (the relation between contrasting concepts, such as *expensive* and *cheap*) and troponymy (a “manner” relation that links verbs like *prance* and *walk*). In addition to relations among synsets, WordNetalso includes lexical relations among specific synset members—morphologically and semantically related form-meanings pairs such as *direct, director*, and *direction*, etc. (Fellbaum and Miller 2003).

WordNet defines membership in a synset as denotational equivalence and substitutability in some, though not all, contexts. But in fact synsets include members that differ along many dimensions, and some are more similar to one another than others. A more subtle representation could label the many ways synset members are related to one another, such as in terms of dialectal variations as in the synsets {*grinder, hero, hoagie, poor boy, submarine*} or register, ranging from formal to taboo words whose use is restricted to particular settings.

Although it was not motivated by Natural Language Processing research, WordNet’s extensive coverage, digital format, and above all its graph structure make it useful for automatic language processing. When WordNet was widely embraced by the English NLP community, wordnets construction began in other languages.

Within the EuroWordNet project (Vossen 1998), lexical databases modeled on the Princeton WordNet were constructed for eight languages. EuroWordNet contributed several fundamental innovations to the wordnet design that have since been adopted by dozens of additional wordnets. One is the definition of a set of Base Concepts, which are characterized by many links to other synsets in wordnets and which are assumed to be universally encoded. Second, to increase the connectivity among synsets, a number of new relations were defined, in particular cross-part-of-speech relations. All relations were marked with a feature value indicating the combinations of relations (conjunctive or disjunctive) and their directionality. Relations may be language-specific rather than apply to all wordnets.

Following EuroWordNet, wordnets were developed for a number of languages around the world. Besides individual efforts, there are wordnets for entire geographic regions, such as BalkaNet (Tufis 2004), African Wordnet (Moropa et al. 2007), Asian wordnet (Robkop et al. 2010) and the Indian wordnets (Sinha et al. 2006). Currently, wordnets exist for some sixty genetically and typologically distinct languages (cf. www.globalwordnet.org).

Many wordnets are created independently and their coverage and design is not uniform. The challenge is to create a framework that accommodates an ever-increasing diversity of languages without shoehorning them into a pattern developed originally for English only. Fellbaum and Vossen (2007) and Vossen and Fellbaum (2009) present an outline of the Global WordNet Grid, a proposed system designed to accommodate the lexicons of all languages and connect them via a language-independent ontology.

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

3.1 Language-specific relations

Synonymy, at least under the somewhat loose definition that WordNet adopts, appears to be a universal phenomenon. And we have not yet encountered a language whose lexicon cannot be organized at least partly by means of hyponymy, meronymy, and antonymy. But some other semantic distinctions are lexicalized in a subset of the world's languages only. This poses challenges for representing the lexemes in a language-independent, universally valid ontology.

3.1.1 Gender lexicalizations

Consider male and female profession nouns like *actor-actress*. While English does not distinguishes gender systematically and nouns like *teacher, lover, friend,* etc. are underspecified with respect to gender, languages like German and Dutch systematically and regularly encode it. The female form is usually derived in a productive fashion by means of a suffix. Because this process is not shared by all languages, only those that mark the distinction will link the noun pairs via a gender relation in their lexicons. The members of such pairs connect to the corresponding classes in the ontology ("male" or "female"); English words are linked to both.

3.1.2 Verbal aspect

Verbal aspect is distinguished and encoded differently across languages. Languages including English and German can differentiate activities from accomplishments by adding a particle to a simplex verb, as in the English pairs *eat up* and *read through* (German *aufessen* and *fertiglesen*). Perfectivity is not obligatorily marked, and a sentence like *Peter read a magazine* is underspecified as to whether or not Peter read the entire magazine, front to back.

Other languages regularly encode semantic distinctions by means of affixes. For example, Slavic languages systematically distinguish between the imperfective, unprefixed and the perfective, prefixed members of a verb pair. Czech has 1,000s of such verb pairs, where one member derived via regular and productive morphology. Do aspectual distinctions belong into the lexicon or the ontology? The fact that they are not universally marked (Romance, for example, uses different conjugational endings but no lexical encoding) might argue for a relation among aspectually related verb pairs in the lexicons of German, English, Czech, etc. (Pala et al. 2008). All verb forms related to the same base form would be linked to one event in the ontology. However, limiting the encoding of verbal aspect to the lexicon and excluding it from the ontology will constrain the reasoning power of the ontology (e.g., if completive *eat up* is not distinguished from progressive *eat*, a system cannot draw any conclusions as to whether the food has been completely consumed or not).

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

3.1.3 Event perspective

Some events involving multiple participants can be expressed with different verbs that profile different participants in the event. For example, converse pairs like *buy* and *sell* express the actions of different participants in the same sale event. FrameNet (Ruppenhofer et al. 2002) captures this difference by referring to distinct Frame Elements—Buyer and Seller—of a single Frame.

While the verbs and the corresponding Agent nouns (*buyer, seller*) each merit their own lexical entries, in the ontology they can be represented so as to reflect different perspectives on the same event. Converse and reciprocal events may be encoded very differently across languages. While English labels the two sides of a sale event with distinct word forms (*buy*, *sell*), others, like German, distinguish them by means of a morpheme (*kaufen* vs. *verkaufen*). 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.

Russian has two different verbs corresponding to English *marry*, depending on whether the grammatical subject refers to the bride or the groom. In such cases, the lexicons need only refer to the event entry in the ontology (*sale*, *marriage*, etc.) and implement equivalence mappings between the terms and lexical entities, leaving the linguistic encoding of distinct verbs and roles to the lexicons of each language.

Crosslinguistic lexicalization patterns show the need for a broader, language-independent treatment that can accommodate all variations on the language level but unifies them on the conceptual, ontological level.

4 Natural language interlingua

Because the lexicons of different languages do not all label the same concepts, a simple mapping from English to the target languages and across the EuroWordNet languages is ruled out in many cases. To interconnect the wordnets, EuroWordNet linked the synsets of each language via an “equivalence relation”, to an interlingual index, or ILI. The ILI permits the mapping of equivalent synsets across all languages connected to the ILI, and thus allows not only for straightforward translations but also for the comparison of the lexicons of different languages both in terms of coverage, relations, and overall lexicalization patterns.

Initially, the ILI was populated with the synsets from the Princeton WordNet, which provided large coverage and was accessible to speakers of the EuroWordNet languages, enabling them to judge semantic equivalence.

EuroWordNet revealed the problems that arise when a natural language becomes the hub connecting the lexicons of other languages. The first concerns *coverage*. No two languages have completely overlapping lexicons. For many concepts, one language may have one or more lexical labels while another language has none. An ILI tied to one specific language clearly reflects only the inventory of the language it is based on, and gaps show up when lexicons of different languages are mapped to it. Using a natural language as the interlingua also may bias the coverage and

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

representation of the wordnets of other languages. Interestingly, those EuroWordNet languages that translated the English WordNet (using the "Expand" method) constructed different wordnets from those that started independently and later mapped onto the ILI (the "Merge" method).

More serious is the question of *equivalence*. The semantic space covered by a word in one language often overlaps only partially with that covered by a similar word in another language, making for less than perfect mappings. An apparently good crosslinguistic match may turn out not to be one when one considers different contexts and social settings. This is the case for connotational differences, tied to specific usages of the words. Second, the mappings among the words and synsets in the ILI may appear to be appropriate on the word level, but there may be a difference in their position within their respective local networks. Such a mismatch necessarily reflects a meaning difference, since in a semantic network the meaning of a node is by definition given in terms of its relations to other nodes. For example, the fact that Dutch lacks a word for "container" does not mean that *bag, box, bottle* etc. do not form a natural category in Dutch, as they do in English by virtue of being children of *container*.

Finally, although WordNet borrows relations like hyponymy and meronymy from ontology, it does not encode the lexicon with such relations in ways that reflect clean *ontological methodology*. As Guarino and Welty (2002a, b) and Gangemi et al. (2003), among others, point out, WordNet's hyponymy relation includes multiple, distinct relations. Earlier versions conflated types, instances, and roles. Thus, *Bill Clinton* was "a type of" *President*, just as *desk* was "a type of" *table*. A later version drew the distinction between Types and Instances, so that proper names referring to people, products, countries, mountains, stars, etc. are now all Instances (Miller and Hristea 2006) and only common nouns can be Types. However, Roles are not presently distinguished from Types, so that *president* and *professor* continue to be represented as "types of" *person* (cf. Sect. 5 for further discussion).

5 From interlingua to ontology

Arguably, using a language-independent interlingua as the hub that connects language-specific lexicons is a better approach to mapping lexicons than a direct mapping. But the interlingua must be able to represent concepts expressed by words in a way that is not biased towards any language or any word-specific linguistic properties at all. The division between words and concepts is reflected in that between the lexicon and ontology.

The use of Princeton WordNet as the interlingua in EuroWordNet blurred this distinction, and the KYOTO project described in Sect. 6 aim to restore it by assigning words on the one hand to wordnet-like structured lexicons and by relegating concepts to ontology.

Lexicons–mappings of labels (words, or lexemes) to concepts (mental representations of entities)—are natural, not products of human reasoning or reflection. They have an internal structure, which is revealed by (often productive) lexicalization

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

patterns and distinct linguistic properties for lexical subclasses (e.g., Levin 1993). But lexicons have many idiosyncracies, such as seemingly unmotivated, "accidental" gaps. Lexicons also show that languages tend to have several labels for given concept (synonymy), though the words may not all be fully equivalent. While the lexicons of all languages may share a core concept-word mapping inventory, language-specific idiosyncracies abound.

WordNet is often called a lexical ontology because it records lexicalized categories and connects them by means of relations familiar from formal ontology. However it differs in significant ways from a formal ontology, an artificially constructed design. Ontologies are language-independent; the linguistic labels in their axioms are merely conveniences and are not to be confused with words used in a natural language. Consequently, the mapping from lexicon to ontology is one from word to concept, rather than across words and languages as in the case of the EuroWordNet ILI. Ontology aims to be completely unambiguous about the meaning of its entries, whereas word meanings are typically fuzzy. Moreover, ontological relations do not necessarily reflect speakers' intuitions about relations among words.

Because each of its entries is unique, clearly defined and distinguished from every other entry. Ontology is preferable over a language-specific lexicon as the hub connecting wordnets of different languages, as argued by Fellbaum and Vossen (2007), Vossen and Fellbaum (2009), and Pease and Fellbaum (2009). This allows for a clean separation between the lexicons and a language-independent, formal representation of the concepts lexicalized by individual wordnets. Moreover, the burden of expressing relations among words and formal concepts can be shared between the lexicons and the ontology. The SUMO ontology (Niles and Pease 2001; 2003) was the first to have been mapped to a number of wordnets and to function as their interlingua.

6 Ontology

In the context of artificial intelligence (AI) and knowledge engineering, ontology is the explicit, formal specification of a conceptualization (Gruber 1992; 1993). 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; furthermore, ontology specifies the relations among concepts (see e.g., Gruber 1993).

The ontology takes input from the lexicons, but on a "selective"' basis, such that not all lexicalized entities are added to the ontology. While the ontology must be able to encode all concepts that can be expressed in any natural language, it need not provide a linguistic encoding—a label—for all words and expressions.

It is desirable that the ontology contain only terms distinguished by essential properties; second, that it be comprehensive and include all distinct concepts that can be represented as Types for all languages; third, that equivalent concepts across languages can be related; fourth, that it allow the definition of all lexicalized

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

concepts having non-essential properties, and finally, that it be logically valid and allow for inferencing.

Guarino and Welty (2002a, b) demonstrated that the WordNet hierarchy, when examined with ontological criteria, can be improved and reduced. Their proposed OntoClean method relies on metaproperties to determine the ontological properties of classes and can be applied to determine the smallest common set of concepts in all languages. The properties of these concepts are *rigidity, essence, dependence* and *unicity*.

Guarino and Welty's rigidity criterion is particularly relevant for the consistent distinction between lexicon and ontology, because languages encode many non-rigid concepts. Rigidity distinguishes Types such as *poodle, Newfoundland, German shepherd* from Roles like *lapdog* and *herding dog*. Types and Roles are not disjunct: a given entity may be both a Type and and a Role at the same time. While a German shepherd will never be a Newfoundland or a poodle, German shepherds may assume different Roles such as that of a herding dog or a lap dog. Only types of dogs are included in the ontology; if a language lexicalizes a role such as *herding dog*, the type hierarchy of the ontology is not extended, but the word is defined in the ontology and marked as a Role (Vossen et al. 1999).[3]

One could include in the ontology all the relations that are found in a semantic network like WordNet. Having done that, the question would be how to express informal linguistic notions with more formal ontological relations. By keeping ontological relation in the formal ontology, and linguistic relations in the lexicon, one can avoid merging two different levels of analysis and yet still capture the information that is needed about both formal concepts and linguistic tokens. An important requirement for the ontology is that it be suitable for automatic reasoning. Therefore, relations in the ontology must be logically consistent and apply strictly.

In a lexicon or a semantic network the meaning of a word can be expressed with natural language definitions. Word meanings as represented in a lexicon are subject to human judgment and introspection. By contrast, in ontology it is solely the axioms as formal statements that gives the terms their meaning. Although the axioms borrow words from natural language, the meanings of these terms are independent of their surface forms. One could replace all the term names with arbitrary unique symbols and they would still have the same meaning. This entails that the meaning of the terms can be tested for consistency with an automated theorem prover, rather than the ontologist having to rely completely on human inspection and judgments of word meaning.

7 Case study: KYOTO, a multilingual information system

KYOTO (Knowledge-Yielding Ontologies for Transition-Based Organization), a project funded by the European Union's Seventh Framework (http://www.kyoto-project.eu), represents the first step toward a Global WordNet. KYOTO rests on

[3] A small number of salient and possibly universally lexicalized roles, including *mother, father, friend* will be included in the type hierarchy.

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

the twin pillars of formal concept representations (ontology) and linguistic representations (lexicons, wordnets), whose division and interrelations allow one to build a domain-specific multilingual wordnet system anchored in a language-independent central ontology. The system is designed to allow easy crosslingual sharing and transfer of information both by automatic systems and by human users without a background in Knowledge Engineering. It enables its users to build crosslinguistic consensus on the meaning and interpretation of language. KYOTO is validated for specific, interlocking domains including biodiversity, climate change and environmental protection (Vossen et al. 2008).

7.1 The KYOTO architecture

KYOTO uses a three-layered knowledge model that separates (1) multilingual general and domain-specific vocabularies linked to (2) multilingual generic and domain-specific wordnets connected to the English WordNet, and (3) a language-independent, formal central ontology, to which all wordnets are linked. Each layer has an internal semantic structure that allows one to connect specific concepts to more general concepts via explicit explicit mapping relations. The ontology contains rich axioms for modeling processes and qualities.

In a first step, human experts identify and specify the locations and sources of domain-relevant documents in different languages. Term extraction from these texts is performed by linguistic miners, so-called term-yielding robots ("tybots"), which identify relevant domain terms and the concepts behind them and relate them to semantic networks (wordnets) in English, Dutch, Spanish, Basque, Italian, Chinese, Japanese. The miners identify possible relations (such as hyponymy) among the members of a phrase or a compound. For example, the miners can suggest that *water* is the polluted entity in the term *water pollution*.

A wiki environment allows ontologically "naïve" users to add domain terms in a way that respects important distinctions among concepts, in particular Rigidity. An editor prompts the domain-experts to identify and encode formal constraints and relations among the terms representing entities, processes and states. This results in a computationally tractable domain ontology that is made available to other user communities where cross lingual validation takes place. The domain wordnets and the ontology are harmonized and anchored to general-coverage wordnets and a generic (domain-independent) ontology.

7.2 The KYOTO ontology

A central question for the system concerns the division of labor between the language-specific lexicons and the ontology (Vossen and Rigau 2010). We outline the criteria for building and distinguishing these two key components of the system.

A top-level ontology is defined as well as a middle level ontology that makes it possible to integrate the environmental knowledge of the applied domain. It would

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

be impossible to represent in the wordnets and in the ontology all complex terms found in domain-specific databases and texts, let alone to attempt automatic inferencing over the terms. Therefore, only a subset of the concepts are represented in the domain-specific wordnets and the generic ontology (which contains only rigid entities) while more specific terms are linked to these via subsumption relations. As a result, the ontology is the direct hub for only a subset of the concepts. In addition, KYOTO makes the assumption that the generic wordnets and vocabularies contain mostly rigid types (e.g., *frog*), whereas domain-specific documents with news and event-specific information typically include in addition non-rigid concepts such as *endangered frogs, endemic frogs* and *alien frogs*. KYOTO allows one to distinguish the rigid entities referred to by a substring of such expressions (e.g., *frog*) and to identify their semantic relation to the states and processes expressed by the remaining constituents (e.g., *endangered*).

A number of mapping relations relate the expressions referring to states and processes in the generic wordnets to the appropriate entries in the ontology.

7.3 Mapping between wordnets and the central ontology

The ontology can represent the processes, states and qualities that are relevant for the KYOTO domain. Mappings were created for highly frequent verbs and adjectives in the domain (e.g., *endanger, endemic*) to these processes, states and qualities in order to differentiate between rigid and non-rigid concepts in the wordnets and to be able to match the non-rigid concepts to the relevant processes. As an example, consider the term *migratory bird*. To reflect that this non-rigid term is a hyponym of *bird* but not a proper subclass, the following mapping was created:

wn:migratory bird sc_domainOf *ont:bird*
wn:migratory bird sc_playRole *ont:done-by*
wn:migratory bird sc_participantOf *ont:migration*

This mapping indicates, first, that the term is used to refer to instances (but not subclasses) of endurants, where the domain is restricted to birds. In addition, the mapping states that the concept in question participates in the process of migration as a participant (in the role of done-by).

The process "migration" is further defined in the ontology, stating that it is an active-change-of-location done-by some endurant, going from a source via a path to some destination. The mapping relations from the wordnet to the ontology need to satisfy the constraints of the ontology, i.e. only roles can be expressed that are compatible with the role-schema of the process in which they participate. The wordnet-to-synset mappings can thus be used to define fairly basic relations relative to the ontology, which represents the full meanings of the terms.

These mappings can clarify many subtle meaning differences among closely related concepts across languages. Consider the following examples:

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

{wn:teacher} English
 →sc_domainOf ***ont:human***
 →sc_playRole ont: ***ont:done-by***
 →sc_participantOf ***ont:teach***
{wn:leraar} Dutch // lit. *male teacher*
 →sc_domainOf ***ont:man***
 →sc_playRole ***ont:done-by***
 →sc_participantOf ***ont:teach***
{wn:lerares} Dutch // lit. *female teacher*
 →sc_domainOf ***ont:woman***
 →sc_playRole ***ont:done-by***
 →sc_participantOf ***ont:teach***

{wn:meat} English
 →sc_domainOf ***ont:cow, sheep, pig***
 →sc_playRole ***ont:patient***
 →sc_participantOf ***ont:eat***
{wn:名 肉, 食物, 餐 } Chinese
 →sc_domainOf ***ont:animal***
 →sc_playRole ***ont:patient***
 →sc_participantOf ***ont:eat***
{wn: غذاء, لحم, طعام} Arabic
 →sc_domainOf ***ont:cow, sheep***
 →sc_playRole ***ont:patient***
 →sc_participantOf ***ont:eat***

On the left, we see mappings for English and Dutch synsets to the role of a *teacher*, where the domain in English is restricted to humans but in Dutch it is differentiated into men and women. On the right, we see representations for edible kinds on meat in English, Chinese and Arabic; note that the domains differ across these languages. The EuroWordNet ILI solution required a mapping from all the non-English synsets to the English ones, blurring often important differences; moreover, it would not allow a flexible representation of non-rigid concepts as in the example above. The solution in KYOTO allows us to keep the differences explicit and at the same time keep the ontology restricted.

7.4 Reasoning and inferencing with KYOTO

The reasoning and inferencing capabilities of KYOTO incorporate the three-layered knowledge model and the notion of an explicit ontology in which a relevant subset of implications is exported to be inserted into knowledge annotation format representations of text.

As the example above shows, classes in the ontology are defined using rich axioms that specify the semantics needed for inferencing: "migration" is represented as an *active-change-of-location done-by* some endurant, going from a source via a path to a destination. At any given point in the ontology development, KYOTO creates an *explicit ontology*, which is a collection of all the implications that apply to a class given the OWL-DL specification of the ontology.

Different surface forms like *migratory birds, bird migration, migration of bids, birds that migrate* are subject to the same ontological implications that build on the relation between the migration process and birds and also provide place holders for other elements in the text to map to the source, path and destination. The same holds for processing of text in languages other than English: regardless of the linguistic (morphosyntactic) structure expressing a concept, the ontology provides the same semantic model.

Journal : 10579 Dispatch : 30-4-2012 Pages : 14
Article No. : 9186 □ LE □ TYPESET
MS Code : LRE881 ☑ CP ☑ DISK

8 Summary and conclusion

There are multiple challenges for aligning wordnets for different languages and create a system that allows crosslinguistic mapping and facilitates automatic language processing. The overall design imposes a clear division between the language-specific lexicons (wordnets) and a formal, language-independent ontology that serves as the hub by which to which all wordnets are interconnected. Ontology is constructed according to strict principles, while the lexicons show a variety of idiosyncracies with respect to the linguistic encoding of concepts and lexical patterns. The KYOTO project provides a framework for the division of labor between ontology and lexicons and for the empirical investigation of the kinds of lexical and sublexical information that ontology can efficiently represent.

Acknowledgments The work described here was funded by the European Union FP7 ICT Work Programme, Project ICT-211423.

References

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fellbaum, C. (Ed.). (2007). *Collocations and idioms: Corpus-based linguistic and lexicographic studies*. Birmingham, UK: Continuum Press.

Fellbaum, C., & Miller, G. A. (2003). Morphosemantic links in WordNet. *Traitement automatique de langue, 44*(2), 69–80.

Fellbaum, C., & Vossen, P. (2007). Connecting the universal to the specific. In T. Ishida, S. R. Fussell & P. T. J. M. Vossen (Eds.), *Intercultural collaboration: First international workshop* (Vol. 4568, pp. 1–16). Lecture Notes in Computer Science, Springer, New York.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2003). Sweetening WordNet with DOLCE. *AI Magazine, 57*(1), 13–24.

Gruber, T. R. (1992). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199–220.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*, 199–220.

Guarino, N., & Welty, C. (2002a). Identity and subsumption. In R. Green, C. Bean, & S. Myaeng (Eds.), *The semantics of relationships: An interdisciplinary perspective*. Dordrecht: Kluwer.

Guarino, N., & Welty, C. (2002b). Evaluating ontological decisions with ontoclean. *Communications of the ACM, 45*(2), 61–65.

Levin, B. (1993). *English verb classes and alternations: A Preliminary investigation*. Chicago, IL: University of Chicago Press.

Miller, G. A. (1990). WordNet. *Special Issue of the International Journal of Lexicography, 3*(4), 151–161.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*, 39–40.

Miller, G. A., & Hristea, F. (2006). WordNet nouns: Classes and instances. *Computational Linguistics, 32*(1), 1–3.

Moropa, K., Bosch, S., & Fellbaum, C. (2007). Introducing the African languages WordNet. In *Proceedings of ALASA*, Pretoria, South Africa.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of FOIS-2* (pp. 2–9). Maine: Ogunquit.

Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the international conference on information and knowledge engineering* (pp. 5–6).

Pala, K., Bosch, S., & Fellbaum, C. (2008). Building resources for African languages. In *Proceedings of the sixth international language resources and evaluation*, Marrakech, Morocco.

Journal : 10579 | Dispatch : 30-4-2012 | Pages : 14
Article No. : 9186 | □ LE | □ TYPESET
MS Code : LRE881 | ☑ CP | ☑ DISK

Pease, A., & Fellbaum, C. (2009). Formal ontology as interlingua. In C. R. Huang & L. Prevot (Eds.), *Ontologies and lexical resources*. Cambridge: Cambridge University Press.

Robkop, K., Thoongsup, S., Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (2010). WNMS: Connceting the distributed Wordnet in the case of Asian WordNet. In *The 5th international conference of the global WordNet association* (*GWC-2010*), Mumbai, India.

Ruppenhofer, J., Baker, C. F., & Fillmore, C. (2002). The FrameNet database and software tools. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the tenth Euralex international congress* (pp. 371–375), Copenhagen, Denmark.

Sinha, M., Reddy, M., & Bhattacharyya, P. (2006). An approach towards construction and application of multilingual Indo–WordNet. In *Proceedings of the third global wordnet conference*, Jeju Island, Korea.

Tufis, D. (Ed.). (2004). The BalkaNet project. *Special Issue of the Romanian Journal of Information Science and Technology, 7*(15), 253–256.

Vossen, P. (Ed.). (1998). *EuroWordNet*. Dordrecht: Kluwer.

Vossen, P., & Fellbaum, C. (2009). Universals and idiosyncrasies in multilingual wordnets. In H. Boas (Ed.), *Multilingual lexical resources*. Berlin: de Gruyter.

Vossen, P., & Rigau, G. (2010). Division of semantic labor in the global wordnet grid. In P. Bhattacharya, C. Fellbaum & P. Vossen (Eds.), *Proceedings of the 5th global WordNet conference*. Narosa Publishing House.

Vossen, P., Peters, W., & Gonzalo, J. (1999). Towards a universal index of meaning. In *Proceedings of ACL-99 workshop, siglex-99, standardizing lexical resources* (pp. 81–90). University of Maryland, College Park, MD.

Vossen, P., Agirre, E., Calzolari, N., Fellbaum, C., Hsieh, S.-K., Huang, C.-R., et al. (2008). Kyoto: A system for mining, structuring, and distributing knowledge across languages and cultures. In *Proceedings of LREC,* Marrakech, Morocco.