Wechsler, Steven 2005. Resultatives under the ‘event-argument homomorphism’ model of telicity. In: N. Erteschik-Shir & T. Rapoport (eds.). *The Syntax of Aspect.* Oxford: Oxford University Press, 255–273.

Wolf, Florian & Edward Gibson 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics* 31, 249–288.

Zaenen, Annie 2006. Mark-up barking up the wrong tree. *Computational Linguistics* 32, 577–580.

*Graham Katz, Washington, DC (USA)*

# 110. Semantics in computational lexicons

## Abstract

*This chapter gives an overview of work on the representation of semantic information in lexicon resources for computational natural language processing (NLP). It starts with a broad overview of the history and state of the art of different types of semantic lexicons in Computational Linguistics, and discusses their main use cases. Section 2 is devoted to questions of how to construct semantic lexicons for Computational Linguistics. We discuss diverse modelling principles for semantic lexicons and methods for their construction, ranging from largely manual resource creation to automated methods for learning lexicons from text, semi-structured or unstructured. Section 3 addresses issues related to the cross-lingual and multi-lingual creation of broad-coverage semantic lexicon resources. Section 4 discusses interoperability, i.e., the combination of lexical (and other) resources describing different meaning aspects. Section 5 concludes with an outlook on future research directions.*

## 1. Representation and computation.

### 1.1. Lexicons in computational semantics.

The development of semantic lexicons in and for computational semantic processing has been shaped by two complementary aspects of lexical meaning. Since words are combined to form complex phrases that express specific meanings, lexical meaning clearly relates to *structural aspects of meaning* in compositional meaning construction, with phenomena such as argument structure, quantifier or adverbial scope, presupposition projection, or anaphoric reference (cf. article 82 (von Stechow) *Syntax and semantics*). But more importantly, the lexicon plays its primary role in the representation of the *lexical meaning of individual words* that build the basis for constructing complex meanings, and that can serve as a basis for recognising and modelling paraphrases, lexically driven entailments, or creative meaning extensions such as metaphor.

Maienborn, von Heusinger and Portner (eds.) 2012, Semantics (HSK 33.3), de Gruyter, 63–93

Computational theories of grammar have been studied extensively in the course of the last decades, with a strong focus on formal modelling and efficient computational processing of compositional meaning construction. Particular focus was put on the design of expressive semantic formalisms, ranging from classical predicate logic to dynamic semantic formalisms, and the design of principled meaning construction methods for diverse grammar frameworks (for an overview see Müller 2010). Since all major computational grammar formalisms are *lexicalised*, it is the computational semantic lexicon in conjunction with compositional meaning construction principles that needs to account for structural semantic phenomena. Phenomena that have received particular attention are quantifier and adverbial scope, plural interpretation, temporal reference, or aspectual properties of events (e.g. Dalrymple et al. 1997, Kamp, van Genabith & Reyle 2011). On the level of representations, a rich body of work is concerned with the compact representation of structural and lexical semantic ambiguities (cf. article 24 (Egg) *Semantic underspecification*).

The meaning representations obtained from computational semantic grammars are typically interpreted using a model-theoretic setting. However, practical uses of computational semantics crucially rely on information about the *lexical meaning of predicates.* As an example, consider a Question Answering system that has to determine that *James Watt was the first to build a working steam engine* is a relevant answer to the query *Who invented the steam engine?*. There are various ways of representing the required lexical semantic knowledge, but none can be considered complete on its own.

In traditional formal semantics, lexical meaning is defined by way of meaning postulates, again interpreted against a model (Carnap 1947), or else by way of lexical meaning relations such as synonymy, antonymy, hyponymy, etc. (Lyons 1977). The semantics of predicate-argument structures describing events or situations has been characterised using semantic or thematic roles, or proto-roles (Fillmore 1976, Dowty 1991). Formal descriptions that define the lexical meaning of predicates have been attempted by way of decompositional analysis (Katz & Fodor 1964). However, agreement on a basic inventory of atomic meaning descriptions has been elusive (Winograd 1978). Most of these approaches to lexical meaning representation have been applied in work on semantic lexicon building for computational grammars at one time or another.

A few proposals also exist for richer semantic characterisations of lexical meaning. Examples include Pustejovsky's generative lexicon that can account for, i.a., the interpretation of metonymy (Pustejovsky 1995), Copestake and Briscoe's work on sense extension (Copestake & Briscoe 1995), or research on the integration of multi-word expressions (Sag et al. 2002). Sharing the concerns of ontological semantics (Nirenburg & Raskin 2004), Cimiano & Reyle (2005) include interfaces to ontological knowledge. Here, the role of ontological knowledge is to provide semantic criteria for ambiguity resolution, and to support inferences on the basis of the derived semantic representations. Finally, substantial research exists on the development of *linking theories* that capture regularities in the syntactic realisation of arguments with specific semantic properties (Bresnan & Zaenen 1990; Grimshaw 1992; Davis & Koenig 2000; Dang, Kipper & Palmer 2000).

All these approaches are mainly concerned with clarifying the formal and computational aspects of representing and processing lexical meaning in computational grammar formalisms, but have not been scaled to large semantic lexicons for broad-coverage, semantically informed NLP systems. Thus, today there exists a good understanding of the mechanisms that are required for the treatment of structural and lexical semantic phenomena in computational grammars – if the information is actually present in the lexicons. The creation of such semantic lexicons – which may involve highly structured representations – is a tight and serious bottleneck.

## 1.2. Standalone semantic lexicons.

The creation of semantic lexicons has been pursued largely independently of computational grammar research. Depending on theoretical assumptions and the intended usage, semantic lexicons are structured according to different aspects of meaning and thus differ considerably in their descriptive devices. Some lexicon accounts characterise the meaning of individual words (or often, their individual *word senses*) by grouping them into *semantic classes* and by defining lexical semantic relations between these classes. Other lexicons try to capture constitutive meaning aspects of lexical items by decomposing their meaning in terms of atomic meaning primitives and define semantic relations between words on the basis of such primitives. Some lexicons, finally, use a combination of these techniques. This section gives an overview of diverse types of semantic lexicons and their modelling principles. We start with lexicons describing the meaning of lexical items in terms of sense definitions, semantic classes, or lexical semantic relations. Argument-taking predicates require in addition semantic descriptions that capture the constitutive meaning relations holding between predicates and their arguments. A number of lexicons is devoted to specific aspects of the meaning of particular word classes (such as nominalisation, factivity, presupposition, or polarity of emotion). Other specialised lexicons focus on the description of the non-compositional semantics of idiomatic expressions, light verbs, or collocations, or relate different modalities. Finally, we address the relation between semantic lexicons and ontologies.

### Lexicons modelling inherent lexical meaning.

Building on influential work in theoretical lexical semantics, in particular Dowty (1979), Jackendoff (1972), Jackendoff (1985) (cf. article 17 (Engelberg) *Frameworks of decomposition*), early attempts to computational lexicon building aimed at providing *inherent meaning descriptions* that can model lexical inferences, the semantic relations between diatheses and paraphrases, or resolve lexical ambiguities in context. Dowty's and Jackendoff's work both aim at inherent, decompositional meaning descriptions in terms of primitive semantic predicates. The aims and scope of decomposition, however, diverge considerably. Dowty's work focuses on explaining systematic meaning relations between diathesis alternations (e.g. inchoative and causative readings of *open* or *close* using primitives like CAUSE and BECOME), and on the ability of these semantic relations to predict the range of possible constructions for different types of predicates. Further aspects concern aspectual properties of verbs. Decomposition is restricted to modelling these grammaticalised categories of lexical meaning, leaving the core lexical semantics of verbs largely unanalysed.

In contrast, Jackendoffs work on Lexical Conceptual Structure (LCS) attempts to capture the lexical meaning of predicates in terms of a set of primitive predicates such as *cause, go* (inspired by physical motion) to define generalisations across predicates (cf. article 19 (Levin & Rappaport Hovav) *Lexical Conceptual Structure*).

Figure 110.1 shows an example for the causative use of *break* with an instrument expressed by a *with*-PP. The heart of the lexicon entry is the semantic description (:LCS) which uses figures to denote semantic categories. It defines the meaning of *break* as "an Agent [1] causes the identity of an Experiencer [2] to become a broken Experiencer [2], using an Instrument [20]" ([9] stands for "Predicate", and [19] for "Instrumental Particle"). The lexicon entry also provides global semantic description of the verb's valency in terms of theta roles (:THETA_ROLES) and selectional restrictions (:VAR_SPEC), as well as mappings to other lexical resources such as WordNet (:WN_SENSE), PropBank (:PROPBANK), and Levin classes (:CLASS).

```
:DEF_WORD "break"

:CLASS "45.1.a"

:WN_SENSE (("1.5" 00787971 00201902)

           ("1.6" 00938146 00231588))

:PROPBANK ("arg0 arg1 arg2(with)")

:THETA_ROLES ((1 "_ag_th,instr(with)"))

:LCS (cause (* thing 1)

       (go ident (* thing 2)

           (toward ident (thing 2)

                         (at ident (thing 2) (break+ed 9))))

       ((* with 19) instr (*head*) (thing 20)))

:VAR_SPEC ((1 (animate +)))
```

Fig. 110.1: LCS lexicon entry for transitive *break* with *with*-PP (Dorr et al. 2001)

Dorr (1997) presents automation techniques to develop LCS-based lexicons, linking LCS representations to Levin classes and WordNet, as seen above. This work proves the applicability of LCS descriptions for special aspects of verb meaning (cf. also VerbNet, below), yet the coverage of LCS-based meaning descriptions is restricted, as is their role in large-scale NLP applications. Pustejovsky (1995) describes the generative capacity of lexical meaning from an opposite viewpoint, assuming a minimal core description and a number of principled operations that allow for systematic sense extensions. The theory considers both verb and noun meanings (cf. article 17 (Engelberg) *Frameworks of decomposition*), but the treatment of nouns in terms of *Qualia Structure* is most widely known: it describes the *constitutive, formal, telic* and *agentive* functions of nouns that account for systematic meaning extensions. While there is no large resource providing qualia information, the CORELEX resource (Buitelaar 1998) models another part of the generative lexicon, namely the systematic polysemy of nouns.

### Lexicons modelling meaning relations.

Another structuring principle for semantic lexicons consists in defining semantic classes, i.e. groups of words that are more or less strictly synonyms, and hierarchical semantic relations among them, in terms of super- and subconcepts. This is the inherent structuring principle underlying taxonomies, which allows us to generalise attributes of concepts at some level in the hierarchy to their subconcepts (cf. article 21 (Cann) *Sense relations*). To account for the pervasive phenomenon of lexical ambiguity, semantic classes need to be distinguished. This may be achieved by way of formal semantic descriptions of the inherent meaning of predicates (see above). Given the difficulty of this task, however, semantic classes are most often defined by way of glosses or textual sense descriptions, combined with linguistic examples.

Early instances of this type of semantic lexicons are machine-readable dictionaries (MRDs) such as the Longman Dictionary of Contemporary English, LDOCE (Procter 1978). LDOCE provides linguistic codes and classifications for word senses, including glosses that use a controlled vocabulary, thus approximating a decompositional analysis. However, MRDs are mostly aimed at human users and contain informal descriptions, inconsistencies, and implicit information. This makes the extraction of general-purpose lexicons from MRDs difficult (Carroll & Grover 1989).

The most widely used resource that adheres to the above-mentioned structuring principles is WordNet, a resource originally motivated by psycholinguistic considerations, and designed for computational as opposed to human usage (Fellbaum 1998). WordNet's semantic classes, called synsets, are defined as groups of *synonymous word senses.* WordNet consists of different hierarchies, one for each major part of speech (noun, verb, adjective, adverb); the main relation between synsets is the *is-a* relation (corresponding to the hyponymy relation between the lexical items in the synsets). Instead of assuming a single top concept, WordNet established 11 top concepts, corresponding to broad conceptual domains (such as entity, event, psychological feature, etc.). Figure 110.2 shows a small excerpt of the WordNet *is-a* hierarchy around the synset for the "automobile" reading of *car.* It shows how WordNet synsets are described with short natural language glosses which are not drawn from a controlled vocabulary, but can nevertheless be viewed as providing a (pre-formal) decompositional analysis. Formalisation of WordNet glosses has been attempted in Mihalcea & Moldovan (2001) by parsing them into logical forms. For many synsets, short example sentences are also available.

Next to hyponymy, WordNet encodes lexical meaning relations such as antonymy, meronymy (part-of relation), entailment, cause, attribute, derivation, etc. These relations provide additional meaning aspects for a given synset, while only indirectly, in terms of their semantic relations to "neighbourhood" concepts (cf. article 21 (Cann) *Sense relations*). The synset *car* from Figure 110.2, for example, is meronymically related to over twenty other synsets, such as *car door, air bag,* or *roof.*

Due to its size (it covers more than 200,000 word senses) and simple structure, WordNet has shown extremely useful in NLP. For example, a wide range of methods exploit its hierarchy to quantify the semantic similarity between words (Budanitsky & Hirst 2006). An unresolved issue, however, is the question of its *granularity.* WordNet uses comparatively fine-grained word senses, which have the potential to convey very specific information. Yet, vagueness and underspecification of lexical meaning in real-world use often make the assignment of a particular sense difficult or impossible (Kilgarriff 1997). Recent

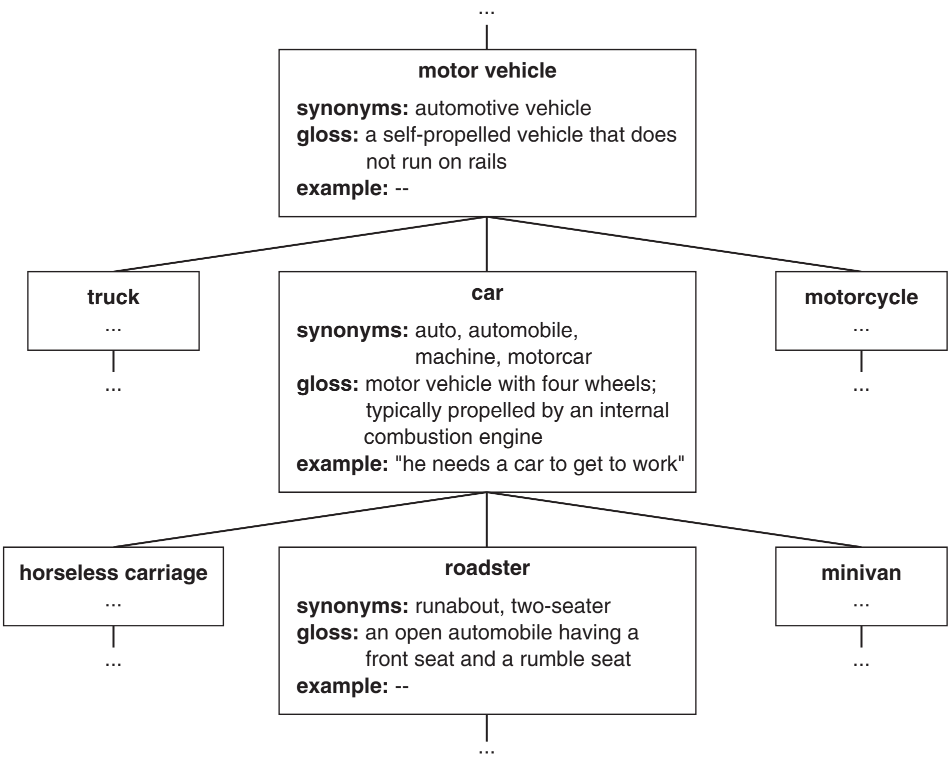

...
motor vehicle
synonyms: automotive vehicle
gloss: a self-propelled vehicle that does not run on rails
example: --
truck
...
...
car
synonyms: auto, automobile, machine, motorcar
gloss: motor vehicle with four wheels; typically propelled by an internal combustion engine
example: "he needs a car to get to work"
motorcycle
...
...
horseless carriage
...
...
roadster
synonyms: runabout, two-seater
gloss: an open automobile having a front seat and a rumble seat
example: --
...
minivan
...
...

Fig. 110.2: WordNet *is-a* hierarchy centered around the primary sense of *car*

work explores strategies to dynamically "group" word senses to find an optimal level of granularity for a given task (Palmer, Dang & Fellbaum 2006).

The increased interest in multimodal applications has lead to the development of multimodal lexicons. Borman, Mihalcea & Tarau (2005) have developed a WordNet-based resource, PicNet, that combines linguistic with pictorial representations for concepts. Such knowledge can be used for better video or image retrieval (Popescu & Grefenstette 2008).

### Lexicons modelling predicate-argument structure.

The characterisation of meaning by way of synonymy, hyponymy and other meaning relations works particularly well for lexical items that refer to entities, as most nouns do. Predicates denoting events or states, such as verbs and deverbal nouns, have a more complex structure in at least two respects: Syntactically, they combine with *arguments*, which requires a semantic characterisation of the arguments in terms of their inherent relation to the event or state (their *semantic role,* such as *agent, patient* or *experiencer*) as well as their linking to surface positions. Also, events and states are often internally structured in terms of aspectual properties. This makes a simple *is-a* hierarchy insufficient to express semantically relevant relations between events and states, for example to draw inferences about the result states of events or the involvement of participants.

A variety of theories exist that characterise the arguments of predicates in terms of thematic or semantic roles, such as AGENT, THEME, LOCATION, etc. (e.g. Gruber 1965, Fillmore 1968, Jackendoff 1972, cf. article 18 (Davis) *Thematic roles*). Classifications were intended to capture syntactic and semantic characteristics of the respective arguments and verb classes. Fillmore (1968), for example, argued for a universal set of atomic thematic roles to capture mainly semantic generalisations, and used these to classify verbs according to the case-frames they allow. Jackendoff (1972) defined a small number of thematic roles in terms of primitive semantic predicates in LCS (see above), and established linking principles to map syntactic and semantic arguments (Jackendoff 1990). However, similar to the general problem of decomposition, no agreement could be reached on a complete and universal set of thematic roles. Dowty (1991) introduced a weaker definition of thematic roles, replacing the set of distinct thematic roles with two "proto-roles" (PROTO-AGENT, PROTO-PATIENT) whose semantics are determined through individual entailments holding for a given predicate and argument. Fillmore (1976) later established a radically different view in introducing Frame Semantics, which assumes concept-specific semantic roles of predicate classes, defined in terms of semantic frames and their frame-specific roles.

A more syntax-oriented view on the semantics of argument structure emerges from the work of Levin (Levin 1993; Levin & Rappaport Hovav 2005). She establishes semantic verb classes ("Levin classes") on the basis of syntactic argument realisations in diathesis alternations. The underlying assumption is that the ability of a verb to occur in certain syntactic alternations is grounded, or a reflex of underlying semantic properties of verbs.

Out of these traditions, a number of large-scale lexicons have emerged that are based on the syntactico-semantic properties of argument-taking predicates, mainly for verbs and mainly for English. Due to the differences between the underlying theories, they differ considerably in their design decisions and structuring mechanisms (Ellsworth et al. 2004; Merlo & van der Plas 2009). Figure 110.3 shows the entries for transitive *break* in the three most widely used predicate-argument structure based lexicons.

*PropBank* (Palmer, Gildea & Kingsbury 2005) is a verb lexicon specifying semantic predicate and role annotations on top of the Penn Treebank, a large English treebank with constituent structure annotation (Marcus, Santorini & Marcinkiewicz 1993); NomBank (Meyers et al. 2004) extends the approach to deverbal nouns. PropBank and NomBank annotate coarse-grained word senses called "rolesets" (like *break.01* in Figure 110.3). The semantic roles ("arguments") are given verb-specific mnemonics (like *breaker*). `Arg0` and `Arg1` correspond to Dowty's proto-agent and proto-patient and thus share meaning across predicates, while arguments with higher numbers are defined in syntactic terms, with limited generalisations. Resources that follow the model of PropBank have been developed for Chinese (Xue 2008) and Korean, although the syntactic nature of PropBank-style roles makes the re-use of English role definitions for other languages difficult.

*VerbNet* (Kipper-Schuler 2005) represents an extension and refinement of Levin verb classes (Levin 1993). It is thus located directly at the boundary between syntax and semantics. On the syntactic side, the lexicon contains syntactic frames (field `Syntax`) with selectional restrictions of verb arguments (field `Roles`). The semantic side is based on intersective Levin classes (Dang et al. 1998), a refinement of Levin's original theory, and defines a hierarchy over verb classes, generally not exceeding a depth of three levels (field `Class`). It assumes a small set of abstract, semantically motivated thematic roles (field

PropBank

```
Roleset  break.01 "break, cause to not be whole"

Verbnet  Class: 1

Roles    Arg0:breaker        Arg1:thing broken

         Arg2:instrument     Arg3:pieces

Example  [Arg0 John] broke [Arg1 the window]

         [Arg2 with a rock].
```

VerbNet

```
Class     45.1

Roles     Agent [+int_control], Patient [+solid], Instrument [+solid]

Syntax    Agent  V    Patient  {with} Instrument

Example   Tony broke the window with   a hammer.

Semantics cause(Agent,E)

          contact(during(E),Instrument,Patient)

          degradation_material_integrity(result(E),Patient)

          physical_form(result(E),Form,Patient)

          use(during(E),Agent,Instrument)
```

FrameNet

```
Frame       Cause_to_fragment

Definition  An Agent suddenly and often violently separates

            the Whole into two or more smaller Pieces,

            resulting in the Whole no longer existing.

Roles       Agent:  The conscious entity, generally a person,

                    that performs the intentional action that

                    results in the Whole being broken into Pieces.

            Cause:  An event which leads to the

                    fragmentation of the Whole.

            Pieces: The fragments of the Whole that

                    result from the Agent's action.

            Whole:  The entity which is destroyed by the Agent

                    and that ends up broken into Pieces.

Example     [AGENT He] *ripped up* [WHOLE the letter].

Semantic Types     Agent=Sentient

Inherits From      Transitive_action

Is Causative of    Fragmentation_scenario

Lexical Units      break.v, break_apart.v, break_down.v,

                   break_up.v, chip.v, cleave.v, ...
```

Fig. 110.3: Lexicon entries for transitive *break* in PropBank, FrameNet, and VerbNet

Roles). For selected meaning aspects, VerbNet provides fine-grained definitions in a decompositional style, using conjunctions of semantic predicates to characterise pre- and post-conditions of the event E as well as temporal and aspectual properties (field Semantics).

*FrameNet* (Fillmore, Johnson & Petruck 2003) is a lexicon in the Frame Semantics paradigm (Fillmore 1976) that groups verbs, nouns, and adjectives into semantic classes (frames) that correspond to abstract situations or events. While a variety of criteria is used in determining frames, most of them tend to be semantic. FrameNet defines semantic roles at the level of individual frames (cf. article 29 (Gawron) *Frame Semantics* for details). Figure 110.3 shows that *break* is analysed as belonging to the Cause_to.Fragment frame, with *definitions* of the frame and roles stated in natural language. Some of the semantic roles are further specified in terms of general semantic types such as Sentient. The frames are organised into a "frame hierarchy" defined by frame-to-frame relations (*inheritance, subframe, causative-of* ect.) that define hierarchical, but also paradigmatic semantic relations, such as successions of events and states in script-like situations. The frame hierarchy also provides mappings between frame-specific semantic roles. Due to its primarily semantics-oriented structuring principles (schematised situations and participant roles), the FrameNet classifications established for English have been successfully transferred to different languages, though not without need for language-specific adjustments in the inventory of frames and roles (Burchardt et al. 2009; Subirats 2009; Ohara et al. 2004).

Due to the differences in their underlying theories, these resources have put different emphasis on syntactic vs. semantic structuring principles, and correspondingly achieve different degrees of generalisations in defining and relating semantic classes. PropBank does not specify relations across lexical items on a formal level, although informal characterisations can be read off the free-text descriptions provided for word senses and role labels. VerbNet achieves a higher degree of generalisation by introducing a certain degree of hierarchical structuring. In addition, it provides strong decompositional semantic definitions of verbs, including thematic roles with selectional preferences. In FrameNet, the coarse-grained semantic classes (frames) typically cover a number of predicates and provide only a limited definition in terms of their sets of semantic roles. Additional characterisations and constraints are only available in free text form. Similar to WordNet, it may be possible to gain considerable information from the hierarchical structure that is defined over frames, in terms of frame-to-frame relations, and in fact this network shows potential for use in NLP tasks (Narayanan & Harabagiu 2004). Nevertheless, the FrameNet resource is still far from complete and requires more rigorous formal definition of frames and frame relations.

A largely unexplored area of semantic lexicon building is the design and creation of lexicons for the difficult classes of non-compositional lexical semantic phenomena. Most computational lexicons assume compositionality in the sense that they specify semantic representations only for "atomic" structures (typically, words), as opposed to idiomatic expressions or multiword expressions. Fellbaum et al. (2006) proposes a model for large-scale lexical resource building focusing on idiomatic expressions coupled with textual data. The SALSA project (Burchardt et al. 2009) investigated special annotation schemes and lexicon entry creation for idiomatic expressions and figurative meanings in the Frame Semantics paradigm.

### Semantic lexicons, ontologies and world knowledge.

There is a close relation between hierarchically structured semantic lexicons such as WordNet and ontologies in that both are organised along the lines of an *is-a* hierarchy. However, there are – at least in theory – two fundamental differences between semantic lexicons and ontologies.

The first distinction lies in the *nature of the objects* that are defined. In semantic lexicons, these are lexical units (words or word senses) of particular languages, while the classes defined in ontologies proper are *concepts* (Gruber 1995) that may or may not be language-independent. This difference becomes obvious once we contrast WordNets for different languages. A comparison of these resources shows that languages can have lexical gaps (cf. the absence of an exact English counterpart to German *Gemü tlichkeit*). At the same time, they may lexicalise distinctions that other languages do not (cf. the English distinction between *isolation* and *insulation* both of which translate into German as *Isolation*). Multilingual semantic lexicons must handle such divergences explicitly. In EuroWordNet, this happens via a *inter-lingual index* (*ILI*) (Vossen 1998).

The second distinction is the *descriptive inventory.* Semantic lexicons categorise lexical items with respect to lexical relations, lexical properties, or predicate-argument structure. In contrast, ontologies provide rigidly defined knowledge-oriented (encyclopedic) relations, attributes and axioms for concepts. For example, the concept *politician* will need to provide typed attributes and relations such as *party*, *period of service*, or *elected by* which are clearly encyclopedic.

In practice, however, the distinction between linguistic meaning and world knowledge is notoriously difficult (Hirst (2004), cf. article 32 (Hobbs) *Word meaning and world knowledge*). Lexical meaning often closely corresponds to conceptual knowledge. Lexical relations like antonymy, synonymy and also entailment are crucially grounded in ontological categories and properties (such as *dead* vs. *alive*, *bachelor* and *unmarried*), which makes them difficult, if not impossible, to distinguish from ontological concepts and relations. Differences show up in cases of linguistic distinctions that do not have an immediate ontological counterpart, as in linguistically conveyed differences in perspectivisation of one and the same event (e.g. *buy* vs. *sell*). On the other hand, some lexical ontologies, such as WordNet, include semantic relations that are truly ontological, such as *part-of*, which adds to terminological confusion.

While linguistic knowledge is often easier to specify than the potentially open-ended field of ontological information, purely linguistic properties are insufficient for NLP applications that require deeper semantic analysis. on the other hand, the knowledge encoded in ontologies cannot be put to use in NLP applications without relating it to the linguistic realisation of the classes and relations. This may be provided in different ways: by constructing an explicit mapping between a semantic lexicon and an ontology (Niles & Pease 2003); by enriching a semantic lexicon with ontological information (Gangemi, Navigli & Velardi 2003), or through construction of hybrid lexicons that include a linguistic and an ontological level of description, such as OntoSem (Nirenburg & Raskin 2004), or HaGenLex (Hartrumpf, Helbig & Osswald 2003).

### Lexicons modelling specific meaning aspects.

A number of lexicon resources concentrate on particular meaning aspects. Some focus on linguistic properties like the implicative behaviour of sentence embedding verbs (Nairn, Karttunen & Condoravdi 2006), the evaluative function of lexical items (Esuli & Sebastian 2006; Pang & Lee 2008), or collocation patterns (Spohr & Heid 2006). Others describe the semantics of particular word classes such as prepositions (Saint-Dizier 2006) or nominalisations (Lapata 2002). Yet other lexicons provide information on generic semantic similarity (Lin 1998) or admissible sentence-level paraphrases (Lin & Pantel 2001). These resources vary widely in how structured they are. On one extreme, they may employ complex graph-based structures (Spohr & Heid 2006), or rest upon in-depth linguistic examination, as in the case of (Nairn, Karttunen & Condoravdi 2006). On the other end of the spectrum, they are sometimes little more than ranked lists of word pairs (Lin 1998).

## 1.3. Semantic lexicons in use.

*From semantic resources to semantic processing.* The various types of knowledge that are represented in computational lexicons are potentially beneficial for a wide range of NLP tasks. We will motivate this claim on a small example from Question Answering, where questions and answer candidates can differ on a number of linguistic dimensions. For example, a potential answer to the question *Whom did Peter see?* may be *The man with the moustache was seen by Peter*, i.e., in passive voice. The relationship between active and passive sentences is best modelled by mapping syntactic (surface) argument positions onto their corresponding *semantic roles*, a process known as semantic role labelling or shallow semantic parsing and pioneered by Gildea & Jurafsky (2002). This is a prime application of predicate-argument structure-based lexicons.

A different problem is posed by *Peter saw the man with his binoculars*, a sentence with an attachment ambiguity where it is unclear whether the binoculars modify the object of the seeing event. Such problems can be addressed by forming semantic classes that describe *selectional preferences* for argument positions, such as the instrument of “see” (Resnik 1996). As illustrated above, some lexicons encode conceptual classes, or selectional restrictions for argument positions.

Next, a sentence like *Peter saw the point of Jack’s argument* should not be considered relevant even though it shares both predicate and subject with the question. The reason is polysemy: here, the sense of “see” can be paraphrased by “understand” while in the question it is closer to “observe”. Selection or assignment of the appropriate word sense in a given context is addressed in the task of *word sense disambiguation* (WSD, Navigli 2009). The by far most widely used sense inventory for this task are the WordNet classes, due to WordNet’s high coverage, and the ability to use the detailed hierarchy to guide generalisation. Finally, some answer candidates can only be recognised as relevant through *inference* (Norvig 1987), such as *The man was identified by the eye witness Peter*: Establishing a relation between this sentence and the question requires the knowledge that being an eye witness necessarily requires an act of observation. This relation might be defined in the inherent meaning of the expression in a lexicon, it might be established through a formal inference process, using knowledge from an ontology, or can be modelled through approximate inference methods (see below).

*Disambiguation in context.* The availability of semantic lexicons and their encoded representations is merely a first step towards their actual use in semantic processing tasks. A serious limitation for their use is that they list the range of possible semantic classes for lexical items, but do not provide specifications as to when these classes are appropriate for a given instance of the lexical item in a specific context. In fact, all applications sketched above crucially depend on *automatic disambiguation* methods that can assign one or more appropriate classes to lexical items in context. Such disambiguation models can be based on a large variety of techniques ranging from knowledge-based or heuristic techniques to statistical models.

Over the last years, robust data-driven methods have been very successful. Such methods can make use of *quantitative information* gained from annotated corpus data which many semantic resource building efforts have produced in parallel with the lexicon resources. The FrameNet database, for example, comes with a large corpus of frame-annotated sentences that were successfully used for training semantic role labelling systems (Gildea & Jurafsky 2002). WordNet provides frequencies of senses through the sense-annotated Semcor corpus (Fellbaum 1998). Data-driven methods for word sense disambiguation are still confronted with serious problems (McCarthy 2009). Due to the highly skewed frequency distribution over senses, supervised models require massive amounts of manual annotations that cannot be achieved on a realistic scale. The performance of unsupervised models, by contrast, is still weak. An alternative to statistical models are knowledge-based methods. The Lesk algorithm (Lesk 1986) and its derivatives compute semantic overlap measures between words in the context of the target word and the words in the sense glosses listed in WordNet for each synset, resulting in a model that is still hard to beat. A promising recent development is the emergence of *knowledge-based* methods that link semantic classes to the vast common-sense knowledge repository Wikipedia, whose articles and link structure can serve as the basis for disambiguation models for the semantic classes without the need for manual annotation (Gabrilovich & Markovitch 2007; Ponzetto & Navigli 2010).

In contrast to independent models for disambiguation, Pustejovsky, Hanks & Rumshisky (2004) propose an integrated *contextual* lexicon model for word senses that associates target entries with syntagmatic patterns of words, so-called *selection contexts*, that determine the assignment of word senses in context.

*Approximate semantic processing*. Semantic analysis in current practical NLP applications is far from comprehensive. This is due to the scarcity of resources on the one hand, and the complexities of fine-grained semantic analysis on the other. Still, currently available resources have been put to use effectively for a variety of semantic analysis tasks that are known to be highly problematic for computational modelling. A commonly used technique is to approach complex phenomena by considering simplified, and thus more tractable, aspects that are accessible to current semantic processing tools and techniques. A crucial factor in the success of this approach is the large amount of *redundancy* in the text collections most NLP tasks are concerned with. Redundancy lowers the requirements on detail and precision, since relevant linguistic material will usually occur more than once. In consequence, even the simple notion of generic *semantic relatedness* is put to use in many applications. It underlies most Information Retrieval systems, and can inform the resolution of syntactic and semantic ambiguities (Dagan, Lee & Pereira 1999; Resnik 1999; Lapata 2002).

With regard to drawing inferences from text, the textual inference framework (Dagan et al. 2009) has risen to prominence. In textual inference, entailment relations between sentences are not defined through a theory of meaning, but rather established by annotator judgments, with the effect of decoupling phenomenon and processing paradigm. A number of approaches have been applied to textual inference, including full-fledged logical inference. However, most approaches are approximate, relying on the partial information present in current semantic lexicons. WordNet, for example, can be used to add hyponym and hyperonym information to analyses; FrameNet and VerbNet can be used to retrieve similar predicates, and to some extent also information about result states (Burchardt et al. 2007).

*Limitations and prospects of current semantic lexicons*. The amount of semantic knowledge encoded in today's semantic lexicons is still limited. As a result, more involved inference problems still remain outside the reach of lexicon-driven approaches. This holds even for the currently most advanced "deep" semantic NLP systems that include large-scale meaning construction and inference machinery, such as Bos (2009). For example, the rejection of the standard interpretation of *Peter and Mary got married* in the context of *Peter married Susan*, *and*

*Mary married John* requires knowledge about the incompatibility of multiple synchronous marriages. One direction of research towards richer semantic resources and processing is the acquisition of such knowledge from corpora, either unstructured text or pre-structured texts from Wikipedia (see Section 2.3.). Another one is the enrichment of semantic representations by building interfaces to manually crafted ontologies such as Cyc or SUMO (Niles & Pease 2003); however, the task of defining flexible interfaces between the lexical and the ontological level is still a challenge (see Section 4.).

## 2. Building semantic lexicons.

Computational lexical semantics has achieved a major break-through in large-scale lexical resource building within the last decade, as evidenced by the resources presented in Section 1. At the same time, current methods are still insufficient to meet the need for deeper analysis, both for general and specialised domains, and, prominently, the need for multilingual resources.

### 2.1. Strategies for building semantic lexicons.

The two main strategies for manual lexicon creation can be seen as opposing poles on a continuum. On one end of the spectrum lie manual resource creation efforts that do not use corpus data at all, relying exclusively on linguistic insight. Great care must be taken not to overlook relevant phenomena, and to achieve a good balance of lexical instances in terms of frequency of occurrence and representativeness of senses.

The other pole is formed by strict corpus-driven lexicon development. This method annotates a corpus from which the lexicon is later extracted. Advantages of this approach include the grounding of the lexicon data in naturally occurring instances, which ensures good coverage of phenomena, and the ability to read quantitative tendencies off the annotations. On the downside, corpus annotation often faces massive redundancy for frequent phenomena. Also, annotation introduces overhead, notably in the effort necessary to guarantee consistency and informativity. Particularly problematic are the ambiguity and vagueness inherent in many semantic phenomena such as word sense (Kilgarriff 1997). Finally, lexicon extraction is confronted with the problem of characterising phenomena across multiple linguistic levels, which requires well-designed interfaces (see Section 4.). In practice, the most feasible strategy for the manual creation of a semantic lexicon is often a compromise. This might involve direct manual creation of the resource that is nevertheless guided by systematic sighting and frequency analysis of the data to encourage high coverage and representativeness. A variety of corpus analysis tools support empirically guided lexicon building through quantitative analysis and linguistically informed search on large corpora: the CQP workbench (Christ et al. 1999), Sketch Engine (Kilgarriff et al. 2004) or the Linguist's Search engine (Resnik & Elkiss 2005). Exemplary corpus annotation can serve to validate analysis decisions and provide data for corpus-driven models.

### 2.2. Conservative methods for data-driven lexicon creation.

Traditional lexicon construction, whether introspective or corpus-driven, proceeds manually and is a long and expensive process. The creation of many semantic lexicons that are in general use, such as WordNet, was only feasible because these resources concentrated on a small set of semantic relations. However, manual lexicon creation strategies can be complemented with semi-automatic methods aimed at extending the coverate of existing lexicon resources. These methods take advantage of corpus-based lexical semantic processing methods and range from simple to challenging.

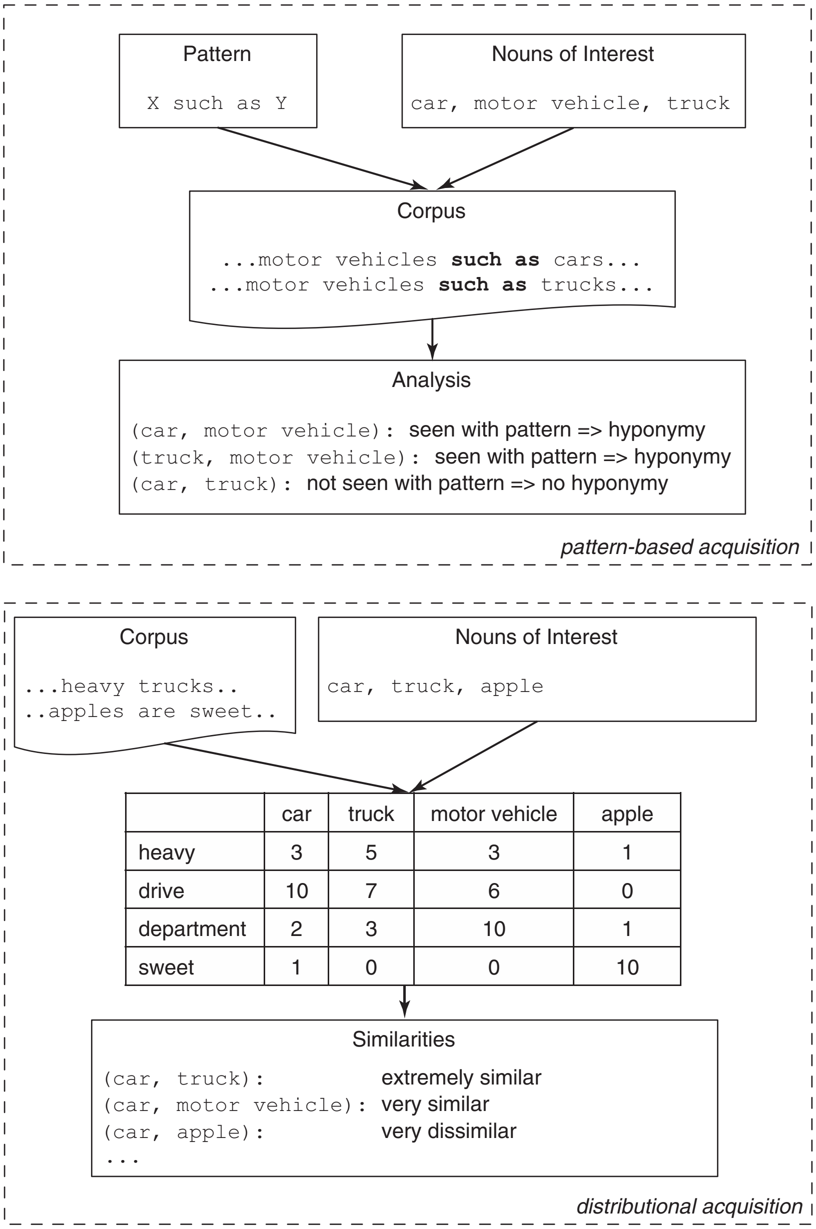
A pressing need that is comparatively simple to address is an increase in coverage to previously unknown lexical items. In *supersense tagging* (Ciaramita & Johnson 2003; Curran 2005) unknown words (usually nouns) are sense-tagged according to a small number of broad WordNet classes. Pennacchiotti & Pantel (2006) build sense vectors characterising synsets that can be used to find the closest WordNet synset for unknown words, bringing together large-scale extraction and integration of semantic relations.

A more challenging goal is the structural extension of a semantic lexicon, which involves shaping new semantic classes or senses, their insertion into the existing lexical hierarchy, and the induction of semantic relations. Fully automated induction of semantic classes, semantic relations, and full ontologies (see below), is still in its infancy. Hence, practical resource creation often reverts to more controlled, semi-automatic methods. For VerbNet, e.g., Korhonen & Briscoe (2004) automatically acquire new Levin classes using corpus-based methods. The integration of this information into the VerbNet hierarchy still requires manual definition of novel semantic classes and predicates, as well as local modifications of the VerbNet hierarchy (Kipper et al. 2006).

### 2.3. Automatic acquisition of semantic lexicons and knowledge bases.

Fully automatic methods try to reduce human effort as completely as possible. As is evident from the previous discussion, completely automatic acquisition is only possible either for coarse-grained classes or by tuning methods to individual relations.

Most such approaches rely on (unannotated) corpora, which are now available for many languages, domains, and genres, often by harvesting from the web. Semantic relations can be gathered from unanalysed corpora by collecting co-occurrence information about words or word pairs, following Harris' (1968) observation that semantically related words tend to occur in similar contexts. Variation in the specification of contexts gives rise to a range of approaches. *Pattern-based* methods use lexico-syntactic templates to identify contexts (typically a small number) that identify individual relations (Hearst 1992). The upper part of Figure 110.4 illustrates this idea for hyponymy relations. In contrast, *distributional* methods record the co-occurrence of individual words with their surrounding context words (e.g., all words within a context window or within a syntactic relationship). Pairwise similarities between the vector representations (e.g., cosine similarity) can then be interpreted as general semantic relatedness (Schütze 1993); see the lower part of Figure 110.4.


Pattern
X such as Y
Nouns of Interest
car, motor vehicle, truck
Corpus
...motor vehicles **such as** cars...
...motor vehicles **such as** trucks...
Analysis
(car, motor vehicle): seen with pattern => hyponymy
(truck, motor vehicle): seen with pattern => hyponymy
(car, truck): not seen with pattern => no hyponymy
*pattern-based acquisition*
Corpus
...heavy trucks..
..apples are sweet..
Nouns of Interest
car, truck, apple
Similarities
(car, truck): extremely similar
(car, motor vehicle): very similar
(car, apple): very dissimilar
...
*distributional acquisition*

| | car | truck | motor vehicle | apple |
|---|---|---|---|---|
| heavy | 3 | 5 | 3 | 1 |
| drive | 10 | 7 | 6 | 0 |
| department | 2 | 3 | 10 | 1 |
| sweet | 1 | 0 | 0 | 10 |

Fig. 110.4: Automatic acquisition of lexical information from corpora

### Learning semantic classes.

Automatic approaches typically start with the induction of semantic classes or senses, i.e., sets of words with similar semantic properties. Unsupervised approaches to this task almost invariably use *clustering* techniques that group words with similar distributional representations into classes (Hindle 1990; Lin 1998; Pantel & Lin 2002). Further examples are Schulte im Walde (2006), who induces verb classes for German purely on the basis of distributional information, and Green, Dorr & Resnik (2004), who induce word classes that are similar in nature to frame-semantic classes by combining evidence from two dictionaries. Prescher, Riezler & Rooth (2000) cluster verb-object pairs to obtain semantic classes. Grenager & Manning (2006) use a structured probabilistic model to induce equivalence classes of arguments across diathesis alternations that resemble PropBank roles.

A major drawback of unsupervised learning methods is that they are incompatible with pre-structuring the domain of semantic classes. This problem is addressed by semi-supervised *bootstrapping approaches.* Here, a small number of initial "seeds" is used to bias the induction of classes towards a desired class structure. Riloff & Jones (1999), Thelen & Riloff (2002) run a pattern-based bootstrapping process to induce semantic classes such as *building*, *human*, *event*, or *weapon.* A major issue in bootstrapping is the acquisition of bad patterns or items, which can "poison" the bootstrapping process. This is usually avoided by confidence-based filtering. In the verbal domain, Miyao & Tsujii (2009) develop a probabilistic supervised model that classifies unseen verbs into the full inventory of VerbNet classes, relying on features extracted from unannotated corpora.

Beyond the level of individual words, surface-oriented acquisition methods may be used to acquire sets of phrases or sentences with similar meanings ("to work for" ⇔ "to be employed by"). This task is called *paraphrase acquisition* and can be based on comparable and parallel corpora (Barzilay & Lee 2003; Bannard & Callison-Burch 2005).

The main challenge in learning semantic classes is the large number of different criteria by which items can be grouped. This is indicated by the large number of classifications proposed in the literature (cf. Section 1.). Consequently, there is no unique "correct" classification, which exposes evaluation against any fixed gold standard to criticism.

### Learning semantic relations.

We now consider the induction of (binary) semantic relations holding between words or semantic classes, the so-called *relation extraction* task. Traditionally, the focus is on nominal relations such as synonymy, hyponymy/hyperonymy (*is-a*) and meronymy (*part-of*) – the relations also found in WordNet. In the pattern-based tradition, Hearst (1992) has used simple surface patterns to induce *is-a* relations. Girju, Badulescu & Moldovan (2006) use a similar approach for meronymy induction. Ruiz-Casado, Alfonseca & Castells (2005) learn extraction patterns and acquire new lexical relations for enriching WordNet using Wikipedia. A recent development is a broader focus on other lexical relations, such as *causation* in work by Pantel & Pennacchiotti (2006). They also use the lexical relations they induce to extend WordNet. Fine-grained relation extraction (Agichtein & Gravano 2000) and classification (Girju et al. 2009) tends to target increasingly encyclopedic relations such as *content-container*, *part-whole*, and thus approaches the domain of ontology learning (see below).

A related task is the *acquisition of inference rules*, which identifies pairs of words where the second follows from the first ("to snore" ⇔ "to sleep"). Such inference rules can be acquired not only on the lexical level, but also for multi-word expressions and phrases (Lin & Pantel 2001; Pantel et al. 2007; Pekar 2008).

Turney & Littman (2005) go beyond the search for individual relations. They develop models to determine the semantic similarity holding between pairs of relation tuples, e.g. *mason:stone – carpenter:wood.* This task extends identification of semantic relations to the task of recognising analogies; it requires representations not only of word meaning, but also of relations between words.

### Learning and populating ontologies.

Techniques for inducing full-fledged ontologies integrate relation learning with class learning, the two tasks described above. It typically begins with the induction of concepts, which may be instantiated with lexical items. The classes are subsequently structured on the basis of semantic relations. These are initially taxonomic, but are subsequently extended by relational and encyclopedic knowledge.

One possibility is to extend the clustering-based methods for inducing semantic classes described above to induce hierarchical structure (Caraballo 1999). Cimiano, Hotho & Staab (2005) refine this technique by using formal concept analysis, using predicate-argument relations as contexts. Unfortunately, the induction of hierarchies with clustering techniques multiplies the problems encountered in analysing and evaluating clustering-induced semantic classes. A promising new development is the injection of global consistency constraints into ontology learning, e.g. by enforcing the transitivity of hyponymy (Snow, Jurafsky & Ng 2006).

Knowledge can also be drawn from other sources. Traditionally, this meant machine-readable dictionaries (Nichols et al. 2006). In the last years, the huge growth of Wikipedia has led to a flurry of work on this resource. Ponzetto & Strube (2007) convert the category structure of Wikipedia into a large *is-a* hierarchy. Ruiz-Casado, Alfonseca & Castells (2005) use Wikipedia as a resource for learning patterns for semantic relations and extend WordNet with newly acquired relation instances. Suchanek, Kasneci & Weikum (2008) construct a large-scale ontology that combines WordNet and Wikipedia. Its taxonomy backbone is formed by WordNet and enriched with facts derived from Wikipedia. While these approaches are able to derive large-scale and high-quality ontological resources (when evaluated against other ontologies, or human judgements), they rely on the existence and correctness of such resources as well as the compatibility of their structuring principles with the target ontology.

Learning semantic knowledge from corpora or structured resources such as Wikipedia currently seems to be the most promising way to solve the acquisition bottleneck. It is, however, inherently restricted to the type of knowledge that is directly or indirectly recoverable from textual or semi-structured resources. General world knowledge remains difficult to acquire from text, as it is often too basic to be conveyed explicitly, even in encyclopedic sources such as Wikipedia.

## 3. Multilingual and cross-lingual aspects.

The development of comprehensive criteria for semantic classification presents itself as a new challenge for each language. Therefore it seems attractive to start from a monolingual model developed for a given language when developing resources for a new language. However, the structure of a monolingual semantic lexicon is not guaranteed to fit other languages, due to conceptual and lexical differences (cf. article 13 (Matthewson) *Methods in cross-linguistic semantics*). In what follows, we discuss strategies for building semantic lexicons for a growing set of languages, and for dealing with cross-linguistic differences in practice.

### 3.1. Manual multilingual resource development.

While some languages (notably English) are fairly well researched, few resources exist for many smaller languages. An important research question is therefore how existing resources in *source* languages (SL) like English can be re-used for efficient development of new *target* languages (TL). Ideally, criteria or even concrete annotation guidelines of the SL can be directly transferred to the TL. This presupposes that the criteria used to structure the SL resource are (at least largely) consistently applicable to other languages. For example, adopting Levin verb classes as structuring principle for a multilingual classification requires that all languages show similar verbal diathesis alternations.

Retaining a tight correspondence between categories and relations across different languages is desirable for another reason: If such correspondences are possible, the design principles evidently capture cross-lingual generalisations. In lexicography, such correspondences allow the study of cross-lingual similarities and differences of lexicalisation patterns (Boas 2005). In NLP, they can be directly exploited for cross-lingual processing, e.g. by translating queries through WordNet synsets or FrameNet classes that relate lexical items across several languages.

The best-known example of parallel lexicon development is WordNet, which has become available for a large number of languages through the EuroWordNet project (Vossen 1998) and the Global WordNet association. Another example is FrameNet, counterparts of which are available or under development for Spanish (Subirats 2009), German (Burchardt et al. 2006), and Japanese (Ohara et al. 2004). PropBank resources are available for Chinese and Korean.

However, the correspondences are rarely, if ever, perfect. There are two strategies how to deal with divergences. EuroWordNet exemplifies *flexible correspondence.* WordNet categories (synsets) are lexical, and therefore tied strongly to individual languages. In EuroWordNet, therefore, resource development in individual languages was largely independent. However, all languages map their synsets onto a so-called "inter-lingual index" (ILI), a unstructured set of language-independent synsets. Specifically, there is an ILI synset for each synset in any of the EuroWordNet languages. However, the links between ILI synsets and synsets of individual languages are not necessarily equivalence (synonymy) links (Peters et al. 1998). For example, the Dutch distinction between "hoofd" (human head) and "kop" (animal head) is mirrored in the existence of two distinct ILI synsets. These ILI synsets are linked to a single English "head" synset by way of a *hyperonymy* link. In this manner, the ILI accommodates structural differences between the individual WordNets.

In contrast, work on FrameNets for new TLs attempts to retain *direct correspondence*, since the categories under consideration, schematised situations, lend themselves more readily to cross-lingual generalisation. Consequently, the structure of FrameNet was used as an initial starting point for most other projects, which restricts the work for new TLs to the assignment of lexical items to pre-defined frames and the collection of examples. Problems arise from FrameNet's assumption that frames are evoked lexically. Figure 110.5 shows an example of a cross-lingual difference in the granularity of lexicalisation. In English FrameNet, the distinction between *driving* a vehicle (as a driver) and *riding* a vehicle (as passenger) was codified in the form of two frames: `Operate_Vehicle` and `Ride_Vehicle`. In German, however, this distinction is not clearly lexicalised: the verb *fahren* can express both situations and cannot be assigned straightforwardly to one of

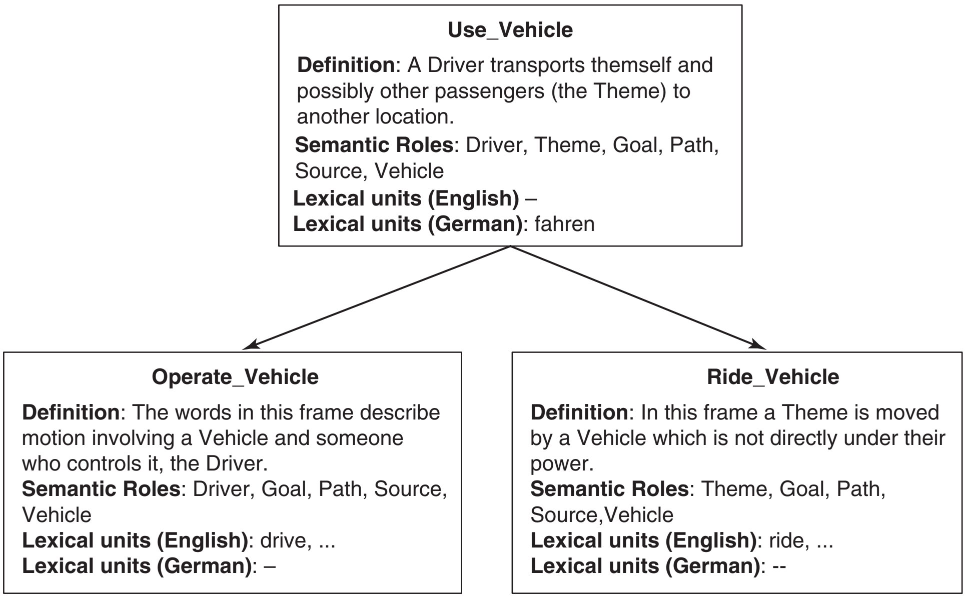

Use_Vehicle
Definition: A Driver transports themself and possibly other passengers (the Theme) to another location.
Semantic Roles: Driver, Theme, Goal, Path, Source, Vehicle
Lexical units (English) –
Lexical units (German): fahren
Operate_Vehicle
Definition: The words in this frame describe motion involving a Vehicle and someone who controls it, the Driver.
Semantic Roles: Driver, Goal, Path, Source, Vehicle
Lexical units (English): drive, ...
Lexical units (German): –
Ride_Vehicle
Definition: In this frame a Theme is moved by a Vehicle which is not directly under their power.
Semantic Roles: Theme, Goal, Path, Source,Vehicle
Lexical units (English): ride, ...
Lexical units (German): --

Fig. 110.5: Example for a cross-lingual divergence (German/English) in FrameNet

the two frames. This situation was resolved by introducing a common superframe, `Use_Vehicle`.

More difficult to amend are general differences in argument realisation patterns between languages (such as differences in the argument incorporation of motion verbs between Romance and Germanic languages). Since in such cases establishing direct correspondence has undesirable consequences, the multilingual FrameNet initiative has decided to revert to an ILI-like flexible mapping when

## 3.2. Cross-lingual resource acquisition.

For many languages, manual resource development is not an option at all. Thus, current research investigates techniques for cross-lingual resource induction to automate this process as completely as possible.

A first straightforward method is to use an existing *bilingual dictionary* to "translate" a SL resource into a TL resource. This method does not have any linguistic context at its disposal, other than the information encoded in the dictionary. Therefore, it requires (i) a high degree of correspondence on the lexical level between the two languages, and (ii) high-quality sense disambiguation for the selection of appropriate translation pairs from the dictionary. Fung & Chen (2004) construct a Chinese frame-semantic predicate classification by mapping English FrameNet entries onto Chinese using two bilingual dictionaries, with a subsequent monolingual disambiguation step, and obtain a high accuracy. While bilingual dictionaries developed for human users are often inconsistent and lack quantitative information, they can also be induced from corpora and used to induce selectional preference information for TLs (Peirsman & Padó 2010).

A second method is the use of *parallel corpora* in a three-step method called *annotation projection* (Yarowsky, Ngai & Wicentowski 2001). In Step 1, the SL side of a parallel corpus is labeled automatically, using the available SL resources. In Step 2, the SL annotations are transferred to the TL, on the basis of automatically induced word alignments. In Step 3, the TL annotations can serve as input either for lexicon creation, as described in Section 2.3., or as training data for new TL labelers. As Resnik (2004) observes, projection can be understood as reducing an unsupervised setting to a supervised setting, in which the TL labels are provided by the SL via word alignment. The validity of the TL labels relies on the so-called "direct correspondence assumption" (Hwa et al. 2002) – namely, that the semantic annotation of a source expression is also valid for its translation in the parallel corpus. This is an issue in particular for structural annotations, such as dependency relations or morphological information, but can be alleviated with filtering. A factor that can greatly affect the quality of target annotations are errors in the word alignments underlying projection. Here, a useful strategy is the exploitation of data redundancy and robust learning methods (Spreyer & Kuhn 2009). Annotation projection has been applied to various semantic phenomena, such as word sense (Bentivogli & Pianta 2005), frame-semantic information (Padó & Lapata 2009), temporal annotation (Spreyer & Frank 2008), or Information Extraction (Riloff, Schafer & Yarowsky 2002).

## 4. Interfaces and interoperability.

The most widely used semantic lexicons in computational semantics concentrate on some well-defined aspect of meaning. For practical purposes, it is therefore often necessary to combine information from several resources. The three most common scenarios are linking semantic lexicons to other levels of description such as syntax or ontologies; the combination of different semantic lexicons; and the combination of general-vocabulary lexicons with domain-specific ones. While these tasks share a number of concerns, such as the compatibility of design principles and granularity issues, each of them poses its own specific challenges.

*Interfaces to morphosyntax.* Using a semantic lexicon for tagging free text with classes or senses usually involves part-of-speech tagging and lemmatisation. Morphological analysis may be required for specific semantic properties, for example tense and aspect for temporal analysis. This step can exploit a large body of work on standardisation (e.g. of tagsets), and divergences between the encodings used in the underlying processors and the coding scheme of a given semantic lexicon are usually easy to resolve. More intricate is the definition of interfaces between syntactic structure and semantic roles in predicate-argument structures. Both symbolic (rule-based) and statistical (feature-driven) interfaces to semantic lexicons need to associate syntactic structures obtained from parsing with their corresponding semantic roles (i.e., linking properties). Currently available parsers deliver constituent- or dependency-based structures, using a wide spectrum of categories and structure-building principles. Therefore, explicit mappings need to be defined between parser output and the syntactic representation used in the lexicon. Here, omission or misclassification of syntactic properties can constitute a serious obstacle for the use of semantic lexicons. Problems of this kind have been addressed in the extraction of lexical resources from PropBank and German FrameNet lexicons from annotated corpora (Babko-Malaya et al. 2006; Burchardt et al. 2008).

*Interfaces to other semantic lexicons*. The coverage of lexical semantic resources that exist for English today is impressive, but when processing free text we are still likely to encounter gaps. This is particularly true for lexicons encoding predicate-argument structure, whose deeper descriptions usually suffer from limited coverage.

This situation has engendered considerable interest in combining and integrating semantic lexicons. Most of the work pursued fully automatic strategies. For example, SemLink (Loper, Yi & Palmer 2007) provides a mapping between VerbNet, PropBank and FrameNet. Often, interest in the mappings is motivated by a particular application: Crouch & King's (2005) Unified Lexicon maps natural language onto a knowledge representation; the goal of Giuglea & Moschitti (2006) and Shi & Mihalcea (2005) is more robust semantic role assignment.

Current approaches rely almost exclusively on simple heuristics to establish inter-resource mappings, such as overlap in verbs between classes, or agreement of verbs on selectional preferences. While the resulting mappings are beneficial for practical purposes, these heuristics cannot deal well with fundamental design differences between resources (such as granularity or the focus on syntactic vs. semantic criteria). Such design differences can be bridged by detailed analysis (Čulo et al. 2008), but appears to be outside the scope of automatic methods.

Interfaces to ontologies. As discussed earlier, semantic lexicons need to be distinguished from ontological resources. Many NLP tasks, however, can benefit from the inclusion of a formal ontology, e.g. as a basis for inference, or as a repository for automatically acquired factual knowledge (as in Information Extraction or Question Answering tasks) (Huang et al. 2010).

An explicit mapping has been manually defined between the English WordNet and the SUMO ontology (Niles & Pease 2003). Mismatches in granularity are covered by explicitly marking non-isomorphic correspondences. A method developed in Spohr (2008) allows to extend this mapping automatically to other languages in EuroWordNet.

Among the largest horizontally and vertically connected resources is the Omega ontology (Philpot, Hovy & Pantel 2010). It integrates the WordNet, VerbNet, FrameNet and LCS lexical resources with a number of upper model ontologies (Hovy et al. 2006). In view of the special needs of NLP applications and given the problems encountered in the alignment of independently developed resources, the OntoNotes project (Pradhan et al. 2007) now undertakes a large *integrated multi-level corpus annotation project* as a basis for corpus-based semantic processing: annotations cover word sense, predicate-argument structure, ontology linking and co-reference relations and are tailored to allow rapid but reliable annotation practice with semi-automatic support for validation.

Interfaces between general and domain-specific resources. The development of NLP applications (e.g., for the natural or social sciences) can involve the creation of domain-specific lexical semantic resources, such as lexicons of medical procedures (Fellbaum, Hahn & Smith 2006) or soccer terms (Schmidt 2006). A major challenge lies in the integration of these specific lexicons with existing generic linguistic resources. Particularly striking are changes in syntactic and semantic properties that can affect general vocabulary items when used in a special domain. Verbs, for example, can show exceptional subcategorisation properties and meanings (e.g. the German *verwandeln (to convert)* with exceptional intransitive use in a soccer context for the special meaning "to turn into a goal"). Similar problems arise at other levels: the use of ontologies requires techniques for interfacing general and domain-specific ontologies. The problem of matching and aligning ontologies automatically is the subject of intensive research in Web Semantics (see article 111 (Buitelaar) *Web Semantics*).

Thus, domain-specific texts require adapted models for parsing, ambiguity resolution as well as special handling in semantic lexicons and their mapping to ontologies. On the other hand, closed domains can also facilitate tasks such as the heuristic selection of word sense (Koeling, McCarthy & Carroll 2005).

*Community efforts for standardisation and interoperability*. In response to such problems, techniques for supporting the standardisation of language resources have been discussed and developed for a considerable time, as in the EAGLES initiative. With developing W3C standards, advanced representation models are being proposed to achieve interoperability across different representation frameworks for language resources (Francopoulo et al. 2006). Recent community efforts work towards ISO-standards for language resources (e.g. in LIRICS). Large community projects are developing resource infrastructures to support interoperability and exchange of resources at a large scale (e.g. CLARIN, FLaReNet; see Calzolari 2008). These projects provide a solid *formal* base for data exchange; agreement on standards for the represented *content* remains a more difficult endeavour.

## 5. Conclusion and outlook.

Natural language processing has seen tremendous achievements in the last decade through the development of a range of large-scale lexical semantic resources. As we have shown, most theoretical frameworks for describing the meaning of words have found their way into lexicon resources, to different degrees and in various combinations.

The creation of WordNet, despite its limitations, can be considered a success story that has engendered stimulating research and advances in semantic processing, comparable to the effect that the Penn Treebank had in the area of syntactic processing. A key role for its feasibility and success was its concentration on a simple relational model of lexical meaning. This allowed rapid development to a sizable resource and offers flexible means for deployment in practical semantic NLP tasks. Its intuitive structure also prepared the ground for developing a multilingual EuroWordNet in a linguistically motivated and computationally transparent architecture. The practical use of such resources is greatly enhanced by the parallel creation of annotated corpora as a basis for induction of automatic annotation and disambiguation components.

Virtually all recent major resource creation efforts, such as FrameNet, PropBank and VerbNet, have adopted the methodological aspects of WordNet and its follow-up projects: (i) concentration on the encoding of a coherent, focused aspect of lexical meaning; (ii) empirical grounding, by using data-driven acquisition methods and providing annotated data for corpus-based learning, and (iii) horizontal multilingual extension, building on experiences gained in 'pilot' languages.

Still, the enormous efforts required for creating more complex lexicons such as VerbNet and FrameNet clearly show that the semantic resource acquisition bottleneck is far from being solved. And while some may still nourish hopes that one day 'the' ultimate, unified semantic theory of the lexicon will be reached, only the tip of the iceberg formed by semantic phenomena has been uncovered.

Largely unexplored is in particular the area of non-compositional lexical semantic phenomena (idioms and support constructions, metaphors) and to what extent they can be integrated with existing semantic lexicons. The situation is similar for the acquisition and integration of lexicons for specific domains. Another issue are fine-grained meaning differences, which are especially important for language generation tasks. These are far from being covered by today's semantic descriptive inventories (Inkpen & Hirst 2006).

Today, we observe three major research directions: (i) the rapid creation of multilingual semantic resources using cross-lingual techniques, capitalising on carefully built existing monolingual resources, (ii) the automated induction of semantic knowledge in monolingual settings, through corpus-based induction methods, and (iii) the integration of complementary semantic lexicons and annotated corpora, both horizontally and vertically, into coherent and interoperable resources.

Statistical, data-driven induction of semantic knowledge is a promising step towards the automation of semantic knowledge acquisition. This area of research is novel and comparatively unexplored, and its methods are faced with the core problems of semantics, in particular the structuring of the semantic space into classes and relations and the identification of salient meaning components. These are challenging decisions even for humans; in addition, corpus-based methods reach their limits when it comes to uncovering deeper aspects of semantic knowledge that cannot be derived from surface phenomena and quantitative analysis. As a result, automatic resource induction is typically used in a semi-automatic fashion that integrates human judgements.

In view of these limitations, novel forms of semantic resource acquisition are being explored that build on collaboratively, human-built resources, folksonomies such as Wikipedia, or specially designed annotation tasks (cf. article 111 (Buitelaar) *Web Semantics*). Structured and unstructured information from Wikipedia can be used for harvesting semantic resources, from taxonomies to ontological attributes and relations. However, Wikipedia's focus is on encyclopedic information rather than lexical semantic information. A new trend builds non-expert contributions for targeted types of knowledge: translation, semantic tagging, etc. using game-like scenarios or Amazon's Mechanical Turk platform.

The move to corpus-based techniques has led to a big momentum and growth in lexical semantic resource building, and approximate methods for using them are well established in natural language processing. But the need for accurate semantic processing persists. More accurate semantic analysis will be needed for tasks that require high precision and that cannot exploit data redundancy. Examples are applications in the areas of knowledge-based natural language understanding and human-machine interaction.

*We thank Rainer Osswald and Alexander Koller for comments and Antonina Werthmann for editorial support.*

## 6. References.

Agichtein, Eugene & Luis Gravano 2000. Snowball: Extracting relations from large plain-text collections. In: *Proceedings of the 5th ACM International Conference on Digital Libraries.* San Antonio, TX: ACM, 85–94.

Babko-Malaya, Olga, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick & Libin Shen 2006. Issues in synchronizing the English Treebank and PropBank. In: *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006.* Sydney: ACL, 70–77.

Bannard, Colin & Chris Callison-Burch 2005. Paraphrasing with bilingual parallel corpora. In: K. Knight, H. T. Ng & K. Oflazer (eds.). *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (= ACL).* Ann Arbor, MI: ACL, 597–604.

Barzilay, Regina & Lillian Lee 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (= HLT-NAACL).* Edmonton, AB: ACL, 16–23.

Bentivogli, Luisa & Emanuele Pianta 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Journal of Natural Language Engineering* 11, 247–261.

Boas, Hans 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography* 18, 445–478.

Borman, Andy, Rada Mihalcea & Paul Tarau 2005. PicNet: Augmenting semantic resources with pictorial representations. In: *Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (= KCVC).* Stanford, CA, 1–7.

Bos, Johan 2009. Applying automated deduction to natural language understanding. *Journal of Applied Logic* 1, 100–112.

Bresnan, Joan & Annie Zaenen 1990. Deep unaccusativity in LFG. In: K. Dziwirek, P. Farrell & E. Meijas-Bikandi (eds.). *Grammatical Relations. A Cross-Theoretical Perspective.* Stanford, CA: CSLI Publications, 45–57.

Budanitsky, Alexander & Graeme Hirst 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics* 32, 13–47.

Buitelaar, Paul 1998. CoreLex: An ontology of systematic polysemous classes. In: N. Guarino (ed.). *Proceedings of the 1st International Conference on Formal Ontology in Information Systems (= FOIS).* Trento: IOS Press, 221–235.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó & Manfred Pinkal 2009. Using FrameNet for the semantic analysis of German: Annotation, representation, and automation. In: H. C. Boas (ed.). *Multilingual FrameNets – Practice and Applications.* Berlin: Mouton de Gruyter, 209–244.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Sebastian Padó & Manfred Pinkal 2006. The SALSA Corpus: A German corpus resource for lexical semantics. In: N. Calzolari et al. (eds.). *Proceedings of the 5th International Conference on Language Resources and Evaluation (= LREC).* Genoa: ELRA-ELDA, 969–974.

Burchardt, Aljoscha, Sebastian Padó, Dennis Spohr, Anette Frank & Ulrich Heid 2008. Constructing integrated corpus and lexicon models for multi-layer annotations in OWL DL. *Linguistic Issues in Language Technology* 1, 1–33.

Burchardt, Aljoscha, Nils Reiter, Stefan Thater & Anette Frank 2007. Semantic approach to textual entailment: System evaluation and task analysis. In: *Proceedings of the 3rd ACL-PASCAL Workshop on Textual Entailment.* Prague: ACL, 10–15.

Calzolari, Nicoletta 2008. Approaches towards a 'Lexical Web': The role of Interoperability. In: *Proceedings of the 1st International Conference on Global Interoperability for Language Resources (= ICGL).* Hong Kong, 34–42.

Caraballo, Sharon A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (= ACL).* College Park, MD: ACL, 120–126.

Carnap, Rudolph 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic.* Chicago, IL: University of Chicago Press.

Carroll, John & Claire Grover 1989. The derivation of a large computational lexicon for English from LDOCE. In: B. Boguraev & T. Briscoe (eds.). *Computational Lexicography for Natural Language Processing.* New York: Longman, 117–133.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann & Esther König 1999. *Corpus Query Processor (CQP). User's Manual.* Stuttgart: IMS, University of Stuttgart.

Ciaramita, Massimiliano & Mark Johnson 2003. Supersense tagging of unknown nouns in WordNet. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (= EMNLP).* Sapporo, 168–175.

Cimiano, Philipp, Andreas Hotho & Steffen Staab 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research* 24, 305–339.

Cimiano, Philipp & Uwe Reyle 2005. Talking about trees, scope and concepts. In: H. Bunt, J. Geertzen & E. Thijsse (eds.). *Proceedings of the 6th International Workshop on Computational Semantics (= IWCS).* Tilburg: ITK, Tilburg University, 90–102.

Copestake, Ann & Ted Briscoe 1995. Semi-productive polysemy and sense extension. *Journal of Semantics* 12, 15–67.

Crouch, Dick & Tracy H. King 2005. Unifying lexical resources. In: *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.* Saarbrücken, 32–37.

Culo, Oliver, Katrin Erk, Sebastian Padó & Sabine Schulte im Walde 2008. Comparing and combining semantic verb classifications. *Journal of Language Resources and Evaluation* 42, 265–291.

Curran, James R. 2005. Supersense tagging of unknown nouns using semantic similarity. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (= ACL).* Ann Arbor, MI: ACL, 26–33.

Dagan, Ido, Bill Dolan, Bernardo Magnini & Dan Roth 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering* 15, i–xvii.

Dagan, Ido, Lillian Lee & Fernando C. N. Pereira 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning* 34, 34–69.

Dalrymple, Mary, John Lamping, Fernando Pereira & Vijay Saraswat 1997. Quantifiers, anaphora, and intensionality. *Journal of Logic*, *Language and Information* 6, 219–273.

Dang, Hoa T., Karin Kipper & Martha Palmer 2000. Integrating compostional semantics into a verb lexicon. In: *Proceedings of the 18th International Conference on Computational Linguistics (= COLING).* Saarbriicken: Morgan Kaufmann, 1011–1015.

Dang, Hoa T., Karin Kipper, Martha Palmer & Joseph Rosenzweig 1998. Investigating regular sense extensions based on intersective Levin classes. In: *Proceedings of the 17th International Conference on Computational Linguistics (= COLING).* Montreal: Morgan Kaufmann, 293–299.

Davis, Anthony & Jean-Pierre Koenig 2000. Linking as constraints on word classes in a hierarchical lexicon. *Language* 76, 56–91.

Dorr, Bonnie 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Journal of Machine Translation* 12, 271–322.

Dorr, Bonnie J., Mari Olsen, Nizar Habash & Scott Thomas 2001. *LCS Verb Database.* College Park, MD: University of Maryland.

Dowty, David 1979. *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ.* Dordrecht: Springer.

Dowty, David 1991. Thematic proto-roles and argument selection. *Language* 67, 547–619.

Ellsworth, Michael, Katrin Erk, Paul Kingsbury & Sebastian Padó 2004. PropBank, SALSA and FrameNet: How design determines product. In: C. Fillmore et al. (eds.). *Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora.* Lisbon, 17–23.

Esuli, Andrea & Fabrizio Sebastian 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In: N. Calzolari et al. (eds.). *Proceedings of the 5th International Conference on Language Resources and Evaluation (= LREC).* Genoa: ELRA-ELDA, 417–422.

Fellbaum, Christane, Udo Hahn & Barry Smith 2006. Towards new information resources for public health – from WordNet to MedicalWordNet. *Journal of Biomedical Informatics* 39, 321–332.

Fellbaum, Christiane (ed.) 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: The MIT Press.

Fellbaum, Christiane, Alexander Geyken, Axel Herold, Fabian Koerner & Gerald Neumann 2006. Corpus-based studies of German idioms and light verbs. *International Journal of Lexicography* 19, 349–360.

Fillmore, Charles J. 1968. The case for case. In: E. Bach & R. T. Harms (eds.). *Universals in Linguistic Theory.* New York: Holt, Rinehart & Winston, 1–88.

Fillmore, Charles J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280, 20–32.

Fillmore, Charles J., Christopher R. Johnson & Miriam R.L. Petruck 2003. Background to FrameNet. *International Journal of Lexicography* 16, 235–250.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet & Claudia Soria 2006. Lexical Markup Framework (LMF*)*. In: N. Calzolari et al. (eds.). *Proceedings of the 5th International Conference on Language Resources and Evaluation (= LREC).* Genoa: ELRA-ELDA, 233–236.

Fung, Pascale & Benfeng Chen 2004. BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. In: *Proceedings of the 20th International Conference on Computational Linguistics (= COLING).* Geneva, 931–935.

Gabrilovich, Evgeniy & Shaul Markovitch 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: M. M. Veloso (ed.). *Proceedings of the 20th International Joint Conference on Artificial Intelligence (= IJCAI).* Hyderabad, 1606–1611.

Gangemi, Aldo, Roberto Navigli & Paola Velardi 2003. The OntoWordNet Project: Extension and axiomatization of conceptual relations in WordNet. In: R. Meersman & Z. Tari (eds.). *Proceedings of On The Move to Meaningful Internet Systems* (= *OTM).* Heidelberg: Springer, 820–838.

Gildea, Daniel & Daniel Jurafsky 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28, 245–288.

Girju, Roxana, Adriana Badulescu & Dan Moldovan 2006. Automatic discovery of part-whole relations. *Computational Linguistics* 32, 83–135.

Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney & Deniz Yuret 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation* 43, 105–121.

Giuglea, Ana-Maria & Alessandro Moschitti 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (= ACL)*. Sydney: ACL, 929–936.

Green, Rebecca, Bonnie Dorr & Philip Resnik 2004. Inducing frame semantic verb classes from WordNet and LDOCE. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (= ACL).* Barcelona: ACL, 375–382.

Grenager, Trond & Christopher D. Manning 2006. Unsupervised discovery of a statistical verb lexicon. In: D. Jurafsky & E. Gaussier (eds.). *Proceedings of the Conference on Empirical Methods in Natural Language Processing(= EMNLP).* Sydney: ACL, 1–8.

Grimshaw, Jane 1992. *Argument Structure.* Cambridge, MA: The MIT Press.

Gruber, Jeffrey S. 1965. *Studies in Lexical Relations.* Ph.D. dissertation. MIT, Cambridge, MA.

Gruber, Thomas R. 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 43, 907–928.

Harris, Zellig S. 1968. *Mathematical Structures of Language.* New York: Interscience Publications.

Hartrumpf, Sven, Hermann Helbig & Rainer Osswald 2003. The semantically based computer lexicon HaGenLex – structure and technological environment. *Traitement Automatique des Langues* 44, 81–105.

Hearst, Marti 1992. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the Fourteenth International Conference on Computational Linguistics (= COLING).* Nantes, 539–545.

Hindle, Donald 1990. Noun classification from predicate-argument structures. In: *Proceedings of the 28th Annual Meeting of the Association for Computational* Linguistics *(= ACL)*. Pittsburgh, PA: ACL, 268–275

Hirst, Graeme 2004. Ontology and the lexicon. In: S. Staab & R. Studer (eds.). *Handbook on Ontologies.* Heidelberg: Springer, 209–229.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel 2006. Onto Notes: The 90% solution. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (= HLT-NAACL).* New York: ACL, 57–60.

Huang, Chu-ren, Nicoletta Calzolari, Aldo Gangemi & Alessandro Lenci (eds.) 2010. *Ontology and the Lexicon: A Natural Language Processing Perspective.* Cambridge: Cambridge University Press.

Hwa, Rebecca, Philip Resnik, Amy Weinberg & Okan Kolak 2002. Evaluating translational correspondance using annotation projection. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (= ACL).* Philadelphia, PA: ACL, 392–399.

Inkpen, Diana & Graeme Hirst 2006. Building and using a lexical knowledge-base of near-synonym differences. *Computational Linguistics* 32, 223–262.

Jackendoff, Ray 1972. *Semantic Interpretation in Generative Grammar.* Cambridge, MA: The MIT Press.

Jackendoff, Ray 1985. *Semantics and Cognition.* Cambridge, MA: The MIT Press.

Jackendoff, Ray 1990. *Semantic Structures.* Cambridge, MA: The MIT Press.

Kamp, Hans, Josef van Genabith & Uwe Reyle 2011. Discourse Representation Theory. In: D. M. Gabbay & F. Guenthner (eds.). *Handbook of Philosophical Logic*, *Vol. 15.* 2nd edn. Dordrecht: Springer, 125–394.

Katz, Jerrold J. & Jerry A. Fodor 1964. The structure of a semantic theory. In: J. J. Katz & J. A. Fodor (eds.). *The Structure of Language: Readings in the Philosophy of Language.* Englewood Cliffs, NJ: Prentice-Hall, 479–518. Originally published in *Language* 39, 1963, 170–210.

Kilgarriff, Adam 1997. I don't believe in word senses. *Computers and the Humanities* 31, 91–113.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell 2004. The Sketch Engine. In: *Proceedings of the 11th EURALEX International Congress.* Lorient, 105–116.

Kipper, Karin, Anna Korhonen, Neville Ryant & Martha Palmer 2006. Extending VerbNet with novel verb classes. In: N. Calzolari et al. (eds.). *Proceedings of the 5th International Conference on Language Resources and Evaluation (= LREC).* Genoa: ELRA-ELDA, 1027–1032.

Kipper-Schuler, Karin 2005. *VerbNet: A Broad-Coverage*, *Comprehensive Verb Lexicon.* Ph.D. dissertation. University of Pennsylvania, Philadelphia, PA.

Koeling, Rob, Diana McCarthy & John Carroll 2005. Domain-specific sense distributions and predominant sense acquisition. In: *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (= HLT/EMNLP).* Vancouver, BC, 419–426.

Korhonen, Anna & Ted Briscoe 2004. Extended lexical-semantic classification of english verbs. In: *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics.* Boston, MA: ACL, 38–45.

Lapata, Mirella 2002. The disambiguation of nominalisations. *Computational Linguistics* 28, 357–388.

Lesk, Michael 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: V. DeBuys (ed.). *Proceedings of the 5th Annual International Conference on Systems Documentation (= SIGDOC).* Toronto, 24–26.

Levin, Beth 1993. *English Verb Classes and Alternations.* Chicago, IL: The University of Chicago Press.

Levin, Beth & Malka Rappaport Hovav 2005. *Argument Realization.* Cambridge: Cambridge University Press.

Lin, Dekang 1998. Automatic retrieval and clustering of similar words. In: *Proceedings of the 17th International Conference on Computational Linguistics (= COLING).* Montreal: Morgan Kaufmann, 768–774.

Lin, Dekang & Patrick Pantel 2001. Discovery of inference rules for question answering. *Journal of Natural Language Engineering* 7, 343–360.

Loper, Edward, Szu-Ting Yi & Martha Palmer 2007. Combining lexical resources: Mapping between PropBank and VerbNet. In: J. Geertzen et al. (eds.). *Proceedings of the 7th International Workshop on Computational Semantics (= IWCS).* Tilburg: ITK, Tilburg University, 118–129.

Lyons, John 1977. *Semantics.* Cambridge: Cambridge University Press.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19, 313–330.

McCarthy, Diana 2009. Word Sense Disambiguation: An overview. *Linguistics and Language Compass* 3, 537–558.

Merlo, Paola & Lonneke van der Plas 2009. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In: *Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (= ACL/AFNLP).* Singapore: ACL, 288–296.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young & Ralph Grishman 2004. Annotating noun argument structure for NomBank. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (= LREC).* Lisbon, 803–806.

Mihalcea, Rada & Dan Moldovan 2001. extended WordNet: Progress report. In: *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources.* Pittsburgh, PA: ACL, 95–100.

Miyao, Yusuke & Jun'ichi Tsujii 2009. Supervised learning of a probabilistic lexicon of verb semantic classes. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (= EMNLP).* Singapore: ACL, 1328–1337.

Müller, Stefan 2010. *Grammatiktheorie.* Tübingen: Stauffenburg.

Nairn, Rowan, Lauri Karttunen & Cleo Condoravdi 2006. Computing relative polarity for textual inference. In: J. Bos & A. Koller (eds.). *Proceedings of the Conference on Inference in Computational Semantics (= ICoS).* Buxton, 67–78.

Narayanan, Srini & Sanda Harabagiu 2004. Question answering based on semantic structures. In: *Proceedings of the 20th International Conference on Computational Linguistics (= COLING).* Geneva, 693–701.

Navigli, Roberto 2009. Word sense disambiguation: A survey. *ACM Computing Surveys* 41, 1–69.

Nichols, Eric, Francis Bond, Takaaki Tanaka, Sanae Fujita & Daniel Flickinger 2006. Robust ontology acquisition from multiple sources. In: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge.* Sydney, 10–17.

Niles, Ian & Adam Pease 2003. Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In: H. R. Arabnia (ed.). *Proceedings of the International Conference on Information and Knowledge Engineering (= IKE).* Las Vegas, NV: CSREA Press, 412–416.

Nirenburg, Sergei & Victor Raskin 2004. *Ontological Semantics.* Cambridge, MA: The MIT Press.

Norvig, Peter 1987. Inference in Text Understanding. In: *Proceedings of the Sixth National Conference on Artificial Intelligence (= AAAI).* Seattle, WA: AAAI Press, 561–565.

Ohara, Kyoko H., Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito & Shun Ishizaki 2004. The Japanese FrameNet Project: An introduction. In: *Proceedings of the LREC Workshop on Building Lexical Resources from Semantically Annotated Corpora.* Lisbon.

Padó, Sebastian & Mirella Lapata 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research* 36, 307–340.

Palmer, Martha, Hoa T. Dang & Christiane Fellbaum 2006. Making fine-grained and coarse-grained sense distinctions both manually and automatically. *Journal of Natural Language Engineering* 13, 137–163.

Palmer, Martha, Dan Gildea & Paul Kingsbury 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31, 71–106.

Pang, Bo & Lillian Lee 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–135.

Pantel, Patrick, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski & Eduard Hovy 2007. ISP: Learning inferential selectional preferences. In: C. L. Sidner et al. (eds.). *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (= HLT-NAACL).* Rochester, NY: ACL, 564–571.

Pantel, Patrick & Dekang Lin 2002. Discovering word senses from text. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (= KDD).* Edmonton, AB: ACM, 613–619.

Pantel, Patrick & Marco Pennacchiotti 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: *Proceedings of the 21st International Conference on Computational Linguistics (= COLING) and the 44th Annual Meeting of the Association for Computational Linguistics (= ACL).* Sydney: ACL, 113–120.

Peirsman, Yves & Sebastian Padó 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (= HLT-NAACL)*. Los Angeles, CA: ACL, 921–929.

Pekar, Viktor 2008. Discovery of event entailment knowledge from text corpora. *Computer Speech & Language* 22, 1–16.

Pennacchiotti, Marco & Patrick Pantel 2006. Ontologizing semantic relations. In: *Proceedings of the 21st International Conference on Computational Linguistics (= COLING) and the 44th Annual Meeting of the Association for Computational Linguistics (= ACL).* Sydney: ACL, 793–800.

Peters, Wim, Piek Vossen, Pedro Diez-Ortas & Geert Adriaens 1998. Cross-linguistic alignment of WordNets with an inter-lingual-index. *Computers and the Humanities* 32, 221–251.

Philpot, Andrew, Eduard Hovy & Patrick Pantel 2010. The Omega Ontology. In: C.-r. Huang et al. (eds.). *Ontology and the Lexicon: A Natural Language Processing Perspective.* Cambridge: Cambridge University Press, 309–322.

Ponzetto, Simone Paolo & Roberto Navigli 2010. Knowledge-rich Word Sense Disambiguation Rivaling Supervised System. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (= ACL).* Uppsala: ACL, 1522–1531.

Ponzetto, Simone Paolo & Michael Strube 2007. Deriving a large scale taxonomy from Wikipedia. In: *Proceedings of the 22nd National Conference on Artificial Intelligence (= AAAI).* Vancouver, BC: AAAI Press, 1440–1445.

Popescu, Adrian & Gregory Grefenstette 2008. A conceptual approach to Web Image Retrieval. In: N. Calzolari et al. (eds.). *Proceedings of the 6th International Conference on Language Resources and Evaluation (= LREC).* Marrakech: ELRA, 28–30.

Pradhan, Sameer, Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel 2007. OntoNotes: A unified relational semantic representation. In: *Proceedings of the 1st IEEE International Conference on Semantic Computing (= ICSC).* Irvine, CA: IEEE Computer Society, 517–526.

Prescher, Detlef, Stefan Riezler & Mats Rooth 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In: *Proceedings of the 18th International Conference on Computational Linguistics (= COLING).* Saarbrücken: Morgan Kaufmann, 649–655.

Procter, Paul (ed.) 1978. *Longman Dictionary of Contemporary English.* New York: Longman.

Pustejovsky, James 1995. *The Generative Lexicon.* Cambridge, MA: The MIT Press.

Pustejovsky, James, Patrick Hanks & Anna Rumshisky 2004. Automated induction of sense in context. In: *Proceedings of the 20th International Conference on Computational Linguistics (= COLING).* Geneva, 924–930.

Resnik, Philip 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61, 127–159.

Resnik, Philip 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11,95–130.

Resnik, Philip 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In: *Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics (= CICLing).* Seoul: Springer, 283–299.

Resnik, Philip & Aaron Elkiss 2005. The linguist's search engine: An overview. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (= ACL).* Ann Arbor, MI: ACL, 33–36.

Riloff, Ellen & Rosie Jones 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In: *Proceedings of the 16th National Conference on Artificial Intelligence (= AAAI) and the 11th Conference on Innovative Applications of Artificial Intelligence (= IAAI).* Menlo Park, CA: AAAI, 474–479.

Riloff, Ellen, Charles Schafer & David Yarowsky 2002. Inducing information extraction systems for new languages via cross-language projection. In: *Proceedings of the 19th International Conference on Computational Linguistics (= COLING).* Taipei, 828–834.

Ruiz-Casado, Maria, Enrique Alfonseca & Pablo Castells 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In: P. S. Szczepaniak & A. Niewiadomski (eds.). *Advances in Web Intelligence.* Heidelberg: Springer, 380–386.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger 2002. Multiword expressions: A pain in the neck for NLP. In: A. F. Gelbukh (ed.). *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (= CICLing).* Mexico City: Springer, 1–15.

Saint-Dizier, Patrick 2006. PrepNet: A multilingual lexical description of prepositions. In: N. Calzolari et al. (eds.). *Proceedings of the 5th International Conference on Language Resources and Evaluation (= LREC).* Genoa: ELRA-ELDA, 877–885.

Schmidt, Thomas 2006. Interfacing lexical and ontological information in a multilingual soccer FrameNet. In: *Proceedings of the 2nd Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies.* Genoa, 75–81.

Schulte im Walde, Sabine 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32, 159–194.

Schutze, Hinrich 1993. Word space. In: S. J. Hanson, J. D. Cowan & C. L. Giles (eds.). *Advances in Neural Information Processing Systems*, *vol. 5.* San Francisco, CA: Morgan Kaufmann, 895–902.

Shi, Lei & Rada Mihalcea 2005. Putting pieces together: combining FrameNet, VerbNet and WordNet for robust semantic parsing. In: A. F. Gelbukh (ed.). *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (= CICLing).* Mexico City: Springer, 100–111.

Snow, Rion, Daniel Jurafsky & Andrew Y. Ng 2006. Semantic taxonomy induction from heterogenous evidence. In: *Proceedings of the 21st International Conference on Computational Linguistics (= COLING) and the 44th Annual Meeting of the Association for Computational Linguistics (= ACL).* Sydney: ACL, 801–808.

Spohr, Dennis 2008. A general methodology for mapping EuroWordNets to the Suggested Upper Merged Ontology. In: N. Calzolari et al. (eds.). *Proceedings of the 6th International Conference on Language Resources and Evaluation (= LREC).* Marrakech: ELRA, 65–72.

Spohr, Dennis & Ulrich Heid 2006. Modelling monolingual and bilingual collocation dictionaries in description logics. In: *Proceedings of the Workshop on Multiword Expressions in a Multilingual Context.* Trentohy, 65–72.

Spreyer, Kathrin & Anette Frank 2008. Projection-based acquisition of a temporal labeller. In: *Proceedings of the 3rd International Joint Conference on Natural Language Processing (= IJCNLP).* Hyderabad, 489–496.

Spreyer, Kathrin & Jonas Kuhn 2009. Data-driven dependency parsing of new languages using incomplete and noisy training data. In: S. Stevenson & X. Carreras (eds.). *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (= CoNLL).* Boulder, CO: ACL, 12–20.

Subirats, Carlos 2009. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In: H. C. Boas (ed.). *Multilingual FrameNets in Computational Lexicography: Methods and Applications.* Berlin: Mouton de Gruyter, 135–162.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum 2008. YAGO: a large ontology from Wikipedia and WordNet. *Journal of Web Semantics* 6, 203–217.

Thelen, Michael & Ellen Riloff 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (= EMNLP).* Philadelphia, PA: ACL, 214–221.

Turney, Peter D. & Michael L. Littman 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning* 60(1–3), 251–278.

Vossen, Piek (ed.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer.

Winograd, Terry 1978. On primitives, prototypes, and other semantic anomalies. In: D. L. Waltz (ed.). *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing (= TINLAP).* Urbana-Champaign, IL: ACL, 25–32.

Xue, Nianwen 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics* 34, 225–255.

Yarowsky, David, Grace Ngai & Roger Wicentowski 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In: *Proceedings of the 1st International Conference on Human Language Technology Research (= HLT).* San Diego, CA, 161–168.

*Anette Frank, Heidelberg (Germany)*
*Sebastian Padó, Heidelberg (Germany)*

# 111. Web semantics

## Abstract

*This article presents an overview of web semantics, i.e., the use and study of semantics in the context of the Web. We differentiate between explicit web semantics, building on Semantic Web standards for web-based knowledge representation (ontologies) and reasoning, and implicit web semantics, building on text and link mining from web resources.*

## 1. Introduction.

This article presents an overview of the emerging field of *web semantics*, divided into *explicit* and *implicit* web semantics.

*Explicit web semantics* is discussed in the context of the Semantic Web, which is fundamentally based on the formal interpretation of web objects (documents, databases, images, etc.) according to an ontology. Web objects are therefore provided with knowledge

Maienborn, von Heusinger and Portner (eds.) 2012, Semantics (HSK 33.3), de Gruyter, 93–106