1. Recent trends in multilingual computational lexicography

Hans C. Boas

1. Introduction

Computational lexicography encompasses the computational methods and tools designed to assist in various lexicographical tasks, including the preparation of lexicographical evidence from many sources, the recording in database form of the relevant linguistic information, the editing of lexicographical entries, and the dissemination of lexicographical products (see Atkins and Zampolli 1994).[1] One of the results of computational lexicography is a dramatic enhancement of Natural Language Processing (NLP) systems through richer machine-readable dictionaries (Boguraev and Briscoe 1989). One early example is the machine-readable version of the Longman Dictionary of Contemporary English (henceforth: LDOCE; Procter 1978), which turned out to be particularly useful for NLP research because it offered detailed subcategorizations of major word classes (see Amsler 1980, Michiels 1982, Ooi 1998, and Fontenelle 2008).

While the emergence of machine-readable dictionaries (MRDs) also facilitated the conception, compilation, and updating of dictionaries for human consumption (Makkai 1980, McNaught 1988), many of the traditional problems of lexicography remained. For example, Atkins (1993: 38) points out that "most machine-readable dictionaries were person-readable dictionaries first." As such, MRDs are often troubled by a variety of problems: omission of explicit statements of essential linguistic facts (Atkins, Kegl, and Levin 1986), unsystematic compiling of one single dictionary, ambiguities within entries, and incompatible compiling across dictionaries (Atkins and Levin 1991). Such problems – as well as new insights – lead lexicographers to revise and restructure MRDs, as, for example, has been

1. For an overview of theoretical and practical aspects of lexicography, see Zgusta (1971), Landau (1989), Béjoint (1994/2001), Svensen (1993), Green (1996), Hartmann and James (1998), Benson (2001), and Fontenelle (2008).

done with the second edition of the LDOCE (Summers 1987) to facilitate its access and use. Despite these issues, MRDs became more widespread during the 1980s, both for human consumption and for machine use. Among the dictionaries made available in machine-readable form were the Collins English Dictionary (1986), the Webster's New World Dictionary (1988), the Oxford Advanced Learner's Dictionary (1989), and the Collins Cobuild English Language Dictionary (1987). Moreover, machine-readable versions of bilingual dictionaries were developed by several publishers, such as the Collins-Robert English-French dictionary (Atkins and Duval 1978). In subsequent years, computational linguists became increasingly interested in developing multilingual lexical resources for a variety of NLP applications, such as machine translation and information extraction.

In this chapter I trace the development of multilingual computational lexicography by covering the period that stretches from the early years to the start of the 21st century. First, I offer a brief account of early machine-readable multilingual lexical resources. In providing this outline, I do not address the many issues raised by theoretical linguistics about the design of mono- and multilingual computational lexical resources (for an overview, see, among others, Atkins and Zampolli 1994, Fontenelle 1997, Heid 1997/2006, Ooi 1998, Calzolari et al. 2001, and Altenberg and Granger 2002). Then, I briefly discuss a number of research initiatives of the 1980s and 1990s that aimed at developing more comprehensive multilingual lexical databases with more semantic information. In this connection, I touch on the increased use of electronic corpora and different theoretical approaches underlying the design of these resources. I next provide an overview of the workflow and design of the FrameNet project, whose outcome, the FrameNet lexical resource for English, forms the basis for the multilingual FrameNets discussed in this volume. Finally, I discuss the development of FrameNets for other languages and compare their design, methods, workflow, tools, and resources used to develop them.

2. The emergence of multilingual lexical databases

The first systematic efforts to produce multilingual MRDs date back to the beginnings of machine translation (MT) in the 1940s when words were organized in lists according to alphabetical order. The source language words were encoded on one side and the target language words on the other side of the lists (see Papegaaij et al. 1986, Ooi 1998). However, this approach proved to be unsuccessful because the translation of words

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 2)

in combination with word-order rules of the target language could not effectively deal with lexical ambiguity. The ensuing range of translations of each potential interpretation of each word resulted in what Ramsay (1991: 30) characterizes as "the generation of text which contained so many options that it was virtually meaningless."

These early exercises in developing MRDs for MT demonstrated the prevalence of the "lexical acquisition bottleneck." To develop large-scale lexical resources for multilingual NLP applications, there were in principle two different approaches: (1) re-using existing resources, or (2) building MRDs from scratch with the help of teams of trained lexicographers. Over the next decades, several efforts were aimed at creating more sophisticated MRDs using these two methodologies. In what follows, I present a brief overview of a select number of these efforts to set up the context for our discussion of the design of multi-lingual FrameNets in sections 4–5.

During the 1950s and 1960s, MRDs became more structured, partially due to the development of more sophisticated syntactic parsing techniques and the newly emerging designs of MT systems that made principled distinctions between linguistic rules, the grammar, and the lexicon (Lehmann 1998). One system that employed such a design was the METAL translation system developed by the Linguistics Research Center at the University of Texas at Austin beginning in the 1960s, whose development continued (with various modifications) until the 1990s (see Slocum 2006). To produce German-to-English translations, the system relied on monolingual dictionaries for English and German that were largely created from scratch, each containing about 10,000 entries. The entries in the METAL dictionary were indexed by canonical form (the usual spelling one finds in a printed dictionary) (Bennett and Slocum 1985). For the input of lexical entries, a lexical default program was developed that allowed the lexicographers to specify only minimal information about a particular entry such as root form and lexical category. The program then heuristically encoded most of the remaining necessary features and values. The METAL lexicon included detailed morpho-syntactic information about part of speech, inflectional class, gender, number, mass vs. count noun, and gradation. With respect to syntax, the lexicon specified the subcategorization frame and the types of auxiliaries. On the semantic side, the METAL lexicon provided only minimal information, namely about the semantic type and the domain (Calzolari et al. 2001: 108–109). The resulting MRD was somewhat limited in scope – it was originally developed for technical translations from German to English – but its minimal entry structure

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 3)

was consistent and provided the types of information needed for the task at hand.

Starting in the early 1980s, the European Community funded a number of multi-lingual NLP projects that relied on MRDs. For instance, the EUROTRA project (Johnson et al. 1985) was aimed at developing a state-of-the-art transfer based MT system for the seven, later nine, official languages of the European Community in order to reduce the amount of time and money spent on the manual translation of documents. In contrast to the older SYSTRAN MT system, which relied heavily on lexical information and only involved minor support for rearranging word order (Gerber and Yang 1997), dictionaries generally played a secondary role in EUROTRA, while grammatical modules were accorded primacy (Alberto and Bennett 1995, Johnson et al. 2003). To keep transfer between languages as simple as possible, operations were reduced to a minimum. In the lexicon, this meant that sense distinctions were identified during the monolingual analysis, while the bilingual resources made use of sense distinctions to relate two lexical entries as translational equivalents. To distinguish different senses, EUROTRA primarily relied on information about argument structure differences, semantic typing of heads, and semantic typing of arguments (see Calzolari et al. 2001: 93). In the following section I discuss various projects that incorporated significantly more semantic information in their multilingual lexical databases than those reviewed above.

3. The focus on semantic information in multilingual lexical databases

During the 1990s, the European Commission explored ways to construct multilingual lexical knowledge bases from machine-readable versions of conventional dictionaries to increase the amount of lexical detail available for multilingual NLP applications at a reasonable cost. To this end, the Research Programs formulated by the Commission made funds available for the ACQUILEX project (Calzolari and Briscoe 1995), which extracted lexical information from multiple MRDs in a multilingual context for English, Dutch, Italian, and Spanish. The goal was the creation of a unique integrated multilingual lexical knowledge base that was maximally re-usable and that was rooted in a common conceptual/semantic structure (Calzolari 1991). This structure was then linked to individual word senses of the languages and was intended to be rich enough to allow for a deep processing model of language (Zampolli 1994). In addition, for each word

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 4)

sense the lexical knowledge base (LKB) contained phonological, morphological, syntactic, and semantic/pragmatic information capable of deployment in the lexical components of a wide variety of practical NLP systems. Figure 1 illustrates the structure of an entry in the LKB.

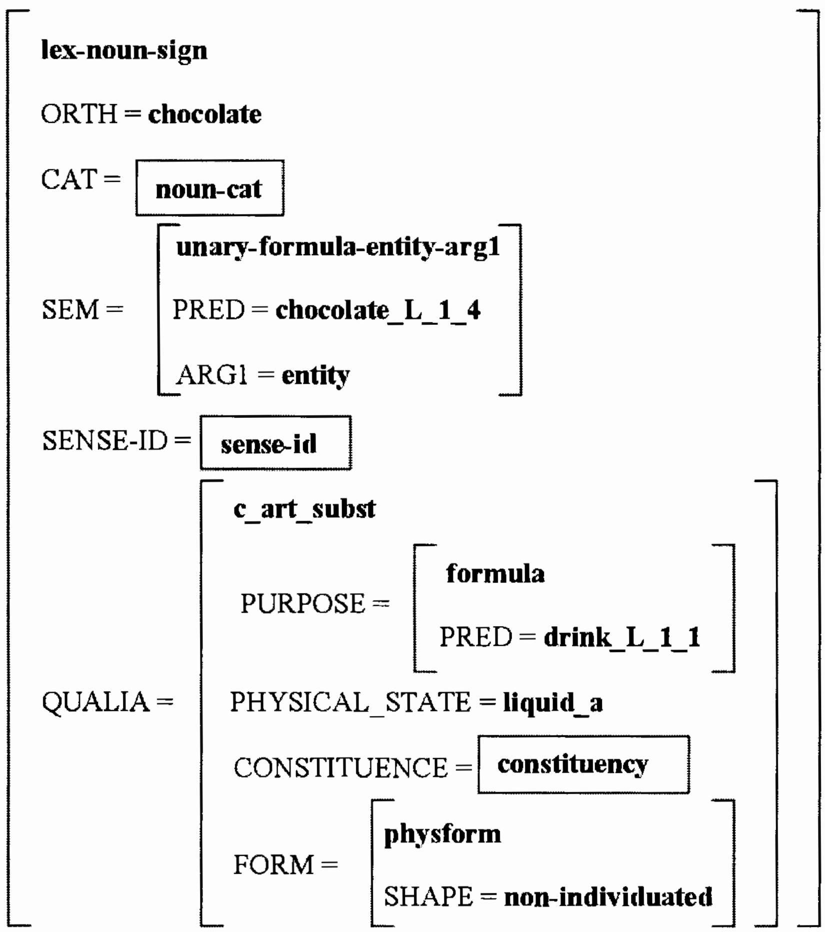

Figure 1. The LKB entry for *chocolate* (Copestake 1992)

Figure 1 shows that more detailed semantic information played an important role in ACQUILEX. Pustejovsky's (1995) concept of "qualia structure" (labeled QUALIA in Fig. 1) served as a theoretical backbone for capturing semantic information and for compiling lexical entries for the project. More specifically, ACQUILEX lexicographers relied on general conceptual templates whose argument slots contain attributes such as agent, set_of, location, used_for, cause_of, color, etc. (for details, see Fontenelle 1997: 13).[2]

Another project funded by the European Commission was EUROTRA-7 (Heid and McNaught 1991), which studied the feasibility of creating large scale shareable and reusable lexical and terminological resources. The project followed up on a 1986 workshop on *Automating the Lexicon: Research and Practice in a Multilingual Environment* (known as the *Grosseto Workshop*), which showed that there was a growing need for standardized and reusable lexical descriptions that could be employed independently of the theoretical framework used for grammatical description (see also Zampolli 1991 and Walker et al. 1995). Focusing on the standards for orthography, phonology, phonetics, morphology, collocation, syntax, semantics, and pragmatics, EUROTRA-7 investigated a broad range of diverse sources of lexical materials as well as different applications relying on lexical components. At the same time the project studied how different theoretical frameworks required various types of information, as well as depth and coverage of descriptions. This investigation resulted in a detailed list of diverging and converging needs, which led to a methodological recommendation for future actions towards developing specifications for reusable linguistic resources. More specifically, the project found that although different theoretical approaches basically described the same facts, they made different generalizations using varying descriptive devices (see Heid et al. 1991).

To provide the various frameworks with reusable lexical and terminological data, EUROTRA-7 recommended going back to the most fine-grained observable differences and phenomena.[3] This methodology would provide extremely detailed linguistic descriptions that would allow the statement of explicit and reproducible criteria for each observable difference. Representing the data in a problem-oriented high-level formalism such as typed feature structures would thus create a common data pool that could form the center of a model consisting of three main areas: acquisition, representation, and application. The recommendations pro-

---

2. For details on the LKB, see Copestake (1992) and Copestake and Sanfilippo (1993).
3. Other projects building on the recommendations of EUROTRA-7 were MULTILEX (MULTILEX 1993), and GENELEX (Antoni-Lay et al. 1994).

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 6)

duced by EUROTRA-7 were significant for the development of future multilingual lexical resources because they explicitly described (1) the initial specifications needed for a model of a reusable lexicon, and (2) the need for standardized formats allowing researchers from academia and industry to use the same lexical resources for a variety of applications, regardless of their theoretical backgrounds.[4]

One of the follow-up projects to EUROTRA-7 was EAGLES (*Expert Advisory Group on Language Engineering Standards*), which started in 1993 with the specific aim to define standards and prepare the ground for future standard provisions. From the outset, EAGLES was not only concerned with standardization of multilingual computational lexicons, but also grammar formalisms, evaluation and assessment, and spoken language. The EAGLES working group on computational lexicons resulted in a series of recommendations for devising standardized architectures for multilingual lexicons.[5] These recommendations were instrumental in the design of the PAROLE-SIMPLE lexicons for twelve European languages (Calzolari et al. 2001: 83), including the semantic lexicons with about 10,000 word meanings. To capture the various dimensions of word meaning, the semantic representation relied on an extension of Pustejovsky's (1995) "qualia structure", which was used as a representational device for expressing the multi-dimensional aspect of word meaning. The semantic layer (SIMPLE) provided a common library of language independent templates, which represented blueprints for any given type to reflect the conditions of well-formedness and to provide constraints for lexical items belonging to that type (Calzolari et al. 2001: 83). The SIMPLE model integrated three types of formal entities, as shown in Figure 2.

The central formal entity was the SemU (semantic unit). It was used to encode word senses as semantic units and could be identified as a semantic type in the ontology, in combination with other types of information that helped to identify a word sense (in addition to distinguish it from other senses of the same lexical item). While SemUs were language specific, those which identified the same sense in different languages were assigned the same semantic type (Calzolari et al. 2001: 83). The second formal entity in the SIMPLE model was the (Semantic) Type, which represented the semantic type assigned to SemUs. The four semantic types were organized

4. See http://www.ilc.cnr.it/EAGLES96/edintro/node11.html.
5. See http://www.ilc.cnr.it/EAGLES96/browse.html#wg2 and http://www.ilc.cnr.it/EAGLES96/EAGLESLE.PDF for details on the recommendations created by EAGLES.

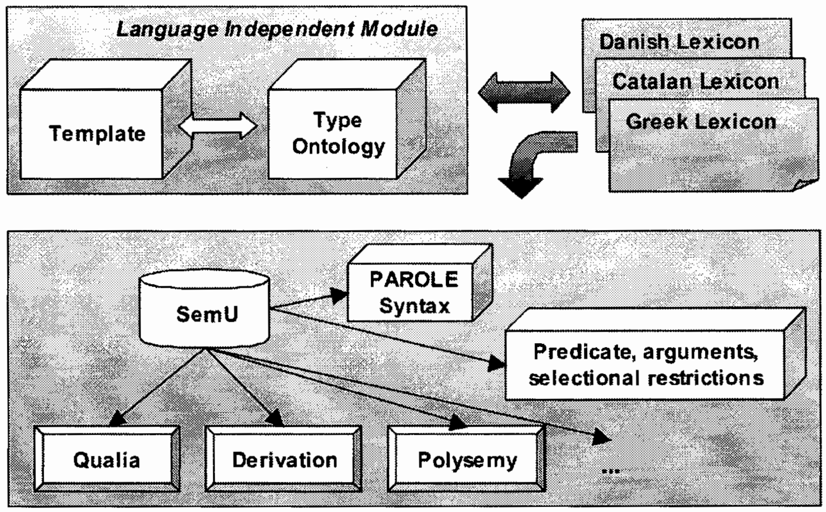

Figure 2. Structure of SIMPLE (Calzolari et al. 2001: 85)

in terms of Pustejovsky's (1995) qualia structures, which in turn were characterized in terms of type-defining information and additional information. The third formal entity was the Template, a schematic structure used by lexicographers to guide, harmonize, and facilitate the encoding of lexical items. The Template stated the semantic type in combination with additional information such as domain, semantic class, gloss, predicative representation, argument structure, polysemous classes, etc. (Calzolari et al. 2001: 83).

The EAGLES initiative and the PAROLE-SIMPLE projects laid much of the groundwork for another initiative for standardizing multilingual lexical resources, namely ISLE (*International Standards for Language Engineering*). One of the outcomes of the ISLE project was a list of detailed suggestions for best practices in the creation and structuring of multilingual lexical entries. At the center of this effort was the MILE (the Multilingual ISLE Lexical Entry), which was envisaged as highly modular and layered. The modularity concept is important in two respects. First, the horizontal level allows independent but linked modules to target different dimensions of lexical entries. Second, the vertical level presumes a layered organization that allows for different degrees of granularity of lexical descriptions, so that both "shallow" and "deep" representations of lexical

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 8)

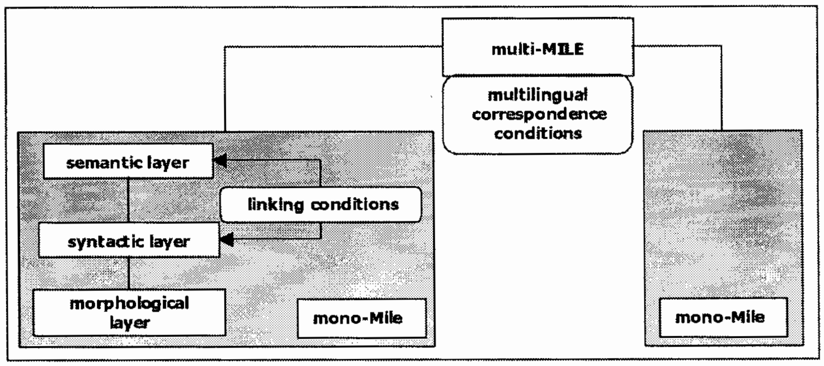

Figure 3. Organization of multi-MILE (Calzolari et al. 2003: 74)

items can be captured. According to the MILE specifications, this feature makes the adoption of different styles and approaches to the lexicon used by existing multilingual systems possible (Calzolari et al. 2003: 8). The organization of MILE, shown in Figure 3, consisted of two modules at the top level, namely mono-MILE, which specified monolingual lexical representations, and multi-MILE, which defined multilingual correspondences. Since space does not permit a full discussion of the MILE (see Calzolari et al. 2003 for full details), consider Figure 3 as an illustration of how each monolingual entry consisted of independent modules providing morphological, syntactic, and semantic information. According to Calzolari et al. (2003: 74), the advantage of this architecture was that it allowed multilingual resource development through the integration of monolingual computational lexicons. This meant that "source and target lexical entries can be linked by exploiting (possibly combined) aspects of their monolingual descriptions."

While the multi-MILE architecture also allowed for the enrichment of syntactic and semantic information that may be lacking in original monolingual lexicons, the authors pointed to a few issues that remained problematic, especially the proper characterization of collocational information and of multi-word expressions. Another important point is the authors' observation that semantic information have "often remained outside standardization initiatives, and nevertheless have a crucial role at the multilingual level" (Calzolari et al. 2003: 74). To lay out the relevant issues surrounding the integration of semantic information in multilingual lexical resources, I now turn to two projects funded by the European

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 9)

Commission that focused on this important task, namely EuroWordNet and DELIS. This overview sets the stage for the discussion in section 3 of how semantic information is encoded in FrameNet, which serves as the basis for the multi-lingual FrameNets discussed in this volume.

During the late 1990s, EuroWordNet (Vossen 1997, Peters et al. 1998) developed a multilingual lexical database connecting independently created WordNets for eight European languages through an unstructured Inter-Lingual-Index (ILI). Each of the individual WordNets was structured along the lines of the original Princeton WordNet for English (Fellbaum 1998), where semantic information is encoded in great detail in the form of lexical semantic relations between synonym sets (the *synsets*, see Miller et al., 1990) such as hyponymy, antonymy, meronymy, etc. (see Cruse 1986). In EuroWordNet, each language-specific WordNet is an autonomous language-specific ontology where each language has its own set of concepts and lexical-semantic relations based on the lexicalization patterns of that language (Vossen 2004).[6] As such, EuroWordNet differentiates between language-specific and language-independent modules. Figure 4 illustrates how a language-independent module, in this case the lexicon of ItalWordNet, is linked to an unstructured ILI and a top concept ontology.

The ILI provides mapping across individual language WordNet structures and consists of a condensed universal index of meaning (1024 fundamental concepts) (Vossen 2001, 2004).[7] Each ILI record consists of a synset and an English gloss specifying its meaning. Although most concepts in each WordNet are ideally related to the closest concepts in the ILI, there are four so-called equivalence relations that map between individual WordNets and the ILI (cf. Vossen 2004: 165–167). Identifying equivalents across languages with EuroWordNet requires a number of steps. One first identifies the correct synset to which the sense of a word belongs in the source language. When there is a one-to-one mapping between synsets and ILI-records, the equivalence relation EQ_SYNONYMY holds

6. In EuroWordNet, there are no concepts for which there are no words or expressions in a language. In contrast, GermaNet (Hamp and Feldweg 1997, Kunze and Lemnitzer 2002), which is a spin-off from the German EuroWordNet consortium, uses non-lexicalized, so-called artificial concepts for creating well-balanced taxonomies.
7. The reason for leaving the ILI unstructured is explained in Vossen et al. (1997: 1) as follows: "A language-independent conceptual system or structure may be represented in an efficient and accurate way but the challenge and difficulty is to achieve such a meta-lexicon, capable of supplying a satisfactory conceptual backbone to all the languages."

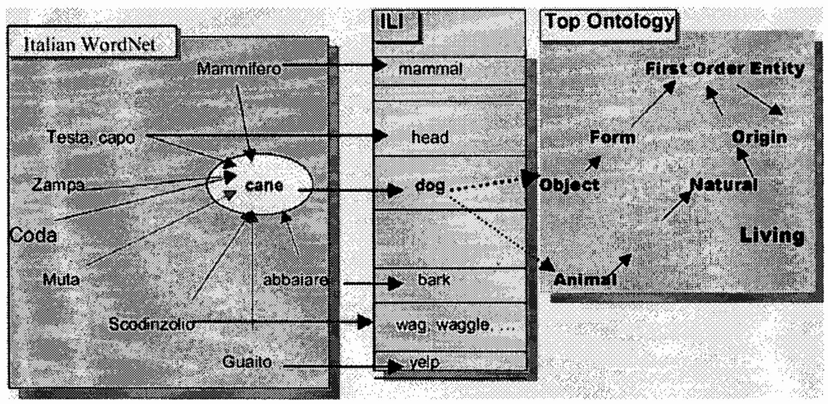

Figure 4. Portion of the ItalWordNet Lexicon for the synset {cane 1} (Calzolari et al. 2003: 23)

and the synset meaning is mapped to the ILI (which is linked to a top-level ontology).

Finally, the corresponding counterpart is identified in the target language by mapping from the ILI to a synset in the target language. The idea behind this mapping relation is described by Vossen et al. (1997: 2) as follows:

> Each synset in the monolingual wordnets will have at least one equivalence relation with a record in this ILI [. . .] Language-specific synsets linked to the same ILI-record should thus be equivalent across languages. The ILI starts off as an unstructured list of WordNet 1.5 synsets, and will grow when new concepts will be added which are not present in WordNet 1.5.

Whenever there is no exact one-to-one mapping that is represented by EQ_SYNONYMY, the mapping is captured by three other mapping relations, which I address only briefly. The first is EQ_NEAR_SYNONYM. It holds when a meaning matches multiple ILI-records simultaneously, when multiple synsets match with the same ILI-record, or when there is some doubt about the precise mapping. The second relation, EQ_HAS_HYPERONYM, holds when a meaning is more specific than any available ILI-record. The third relation is EQ_HAS-HYPONYM. It holds when a meaning can only be linked to more specific ILI-records (for details see Vossen (2004: 165)).

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 11)

The level of detail with which EuroWordNet approached lexical semantic relations in individual languages (as well as cross-linguistically) is remarkable. Its success is reflected by the fact that a number of follow-up projects adopted this approach, such as GermaNet for German (Kunze and Lemnitzer 2002) and a number of projects under the auspices of the Global WordNet Association.[8] The current move towards a Global WordNet Grid (GWG) (Vossen and Fellbaum, this volume) seeking to link WordNets of an even greater variety of languages with each other represents a further step towards providing more semantic information in multilingual lexical databases.

Another project seeking to incorporate more semantic information in multilingual lexical databases was the corpus-based DELIS project (Emele and Heid 1994).[9] Unlike other projects, DELIS focused on the problems of lexicographic relevance and worked towards developing tools that allowed lexicographers to efficiently access corpus materials for specific descriptive tasks (see Heid 1996b). To determine the feasibility of such a corpus-based approach, DELIS developed a set of parallel monolingual lexicon fragments for English, French, Italian, Danish, and Dutch. The lexicon fragments were parallel in that (1) they covered the same fragment (the most general verbs of sensory perception and of speech), and (2) they were based on the same theoretical approaches and on comparable classifications and descriptive devices (Heid 1996a). Using a typed feature structure system (Emele 1993), DELIS also aimed at systematically comparing and describing the interaction between syntax and semantics in the five languages. On the syntactic side, DELIS adopted a syntactic description close to that of Head-Driven Phrase Structure Grammar (Pollard and Sag 1994). On the semantic side, DELIS described lexical items in terms of Frame Semantics (see Fillmore (1985) and section 3). The dictionary architecture in DELIS exhibited three distinct characteristics. The first was that the DELIS architecture was modular. There were separate hierarchical modules for each of the descriptive levels encoded, i.e. Morphosyntax, Syntax, and Semantics (see Heid 1996a: 296).

As Table 1 illustrates, the levels included predicate-argument structures with semantic roles, a description of subcategorized elements in terms of

8. See http://www.globalwordnet.org/gwa/wordnet_table.htm for a list of language-specific WordNet projects.
9. DELIS (Descriptive Lexical Specifications and Tools for Corpus-based Lexicon building) was funded in part by the European Union and operated from February 1993 through April 1995.

grammatical functions, and a description of the phrase structural constructs through which the arguments are realized. One advantage of this approach was that the interaction between the levels could be expressed by means of relational statements, effectively implementing linking rules. This was possible because for each level-specific module there was an inventory of descriptive devices such as a role inventory, an inventory of grammatical functions, and an inventory of phrase types. Another advantage was that individual monolingual lexicons were modules which could be combined to form a multilingual lexicon (Heid 1996b).

Table 1. Summary of components and classes (Heid 1996b)

Construct → Level ↓	Descriptive Devices	Constellations (Classes)
lexical semantics	ROLES	ROLE CONSTELLATIONS
functional syntax	GRAMM. FUNCTIONS	TOPMOST SYNTACTIC CLASSES
categorial syntax	SYNTACTIC CATEGORIES, PHRASE TYPES	SPECIFIC SYNTACTIC CLASSES

The second defining characteristic was that DELIS dictionaries were classificatory in that the description of each level was organized in monotonic multiple inheritance hierarchies of types, each type defining a class of linguistic objects from a particular point of view. This approach allowed DELIS lexicographers to define for a lexical semantic field the combinations of semantic roles, in combination with a syntactic subcategorization hierarchy (Heid 1996a).

The third central feature of DELIS was that there was neutral access to different types of lexical information. This meant that for a given lexical entry, information was flowing together from different descriptive levels without privileging any single level, thereby guaranteeing access neutrality (Heid 1996a). As Figure 5 illustrates, each descriptive level is a separate, usually hierarchical component of the lexical specifications. This means that single readings (indicated by a black dot in Figure 5) inherit from the relevant classes of each component (Heid 1996b).

To illustrate the structure of a DELIS entry, consider Figure 6, which represents the schema of a verb entry in the DELIS dictionary. The top section of the entry ("LEMMA") specifies the head form of the lemma. The mid-section of the entry encodes Frame Element Groups (FEGs), which combine the description of the participants (in terms of semantic

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 13)

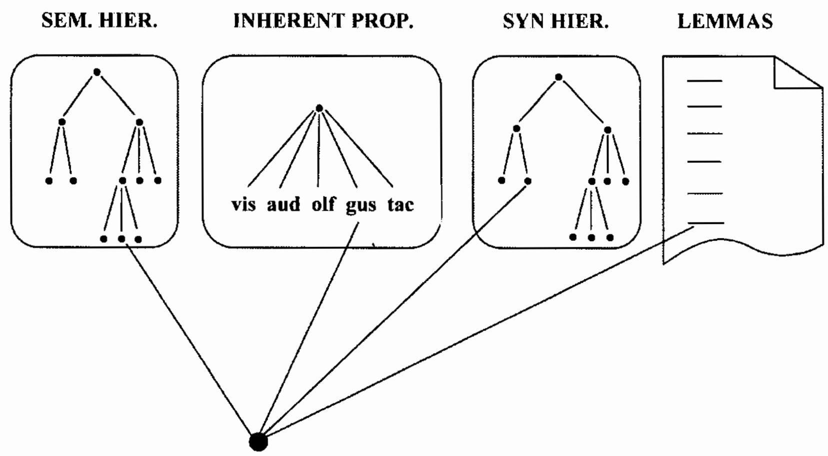

Figure 5. Access-neutrality: information from different levels flowing together, no single level privileged (Heid 1996a)

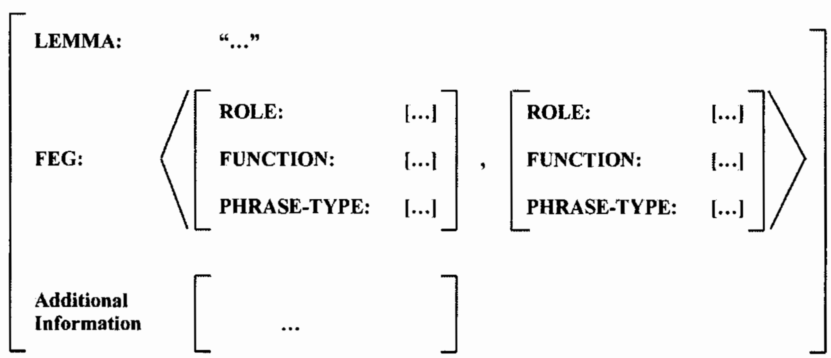

Figure 6. Schema of a verb entry in the DELIS dictionary (Heid 1996a)

roles, cf. Fillmore 1985) with a syntactic description in terms of grammatical functions (subject, direct object, etc.) and syntactic categories (Heid 1996b).

As I will show in the remainder of this chapter, the DELIS architecture is of particular interest because it implemented a number of design features that later became important for the English FrameNet project, which began its work two years after DELIS came to an end. More importantly, however, is the fact that DELIS laid much of the conceptual

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 14)

groundwork for the design of multilingual FrameNets (see also Heid 1997), which are the topics of the papers in this volume.

4. The emergence of multilingual lexical databases

The FrameNet project builds on Frame Semantics, a theory developed by Charles Fillmore and his associates over the past three decades. It differs from other theories of lexical meaning in that it builds on common backgrounds of knowledge (semantic "frames") against which the meanings of words are interpreted.[10] A "frame is a cognitive structuring device, parts of which are indexed by words associated with it and used in the service of understanding" (Petruck 1996: 2). The central concepts underlying Frame Semantics are characterized by Fillmore and Atkins (1992: 76–77) as follows.

> A word's meaning can be understood only with reference to a structured background of experiences, beliefs, or practices, constituting a kind of conceptual prerequisite for understanding the meaning. Speakers can be said to know the meaning of the word only by first understanding the background frames that motivate the concept that the word encodes. Within such an approach, words or word senses are not related to each other directly, word to word, but only by way of their links to common background frames and indications of the manner in which their meanings highlight particular elements of such frames.

Consider, for instance, the `Compliance` frame, which is evoked by several semantically related words such as *adhere*, *adherence*, *comply*, *compliant*, and *violate*, among others (Johnson et al. 2003). The `Compliance` frame represents a kind of situation in which different types of relationships hold between "Frame Elements" (FEs), which are defined as situation-specific semantic roles.[11] This frame concerns ACTS and STATES_

10. For an overview of Frame Semantics, see Fillmore (1970, 1975, 1976, 1977a, 1977b, 1982, 1985), and Fillmore and Atkins (1992, 1994, 2000), among others. Furthermore, the September 2003 issue of the *International Journal of Lexicography* was devoted exclusively to FrameNet.
11. Names of Frame Elements (FEs) are capitalized. Frame Elements differ from traditional universal semantic (or thematic) roles such as Agent or Patient in that they are specific to the frame in which they are used to describe participants in certain types of scenarios. "Tgt" stands for target word, which is the word that evokes the semantic frame.

OF_AFFAIRS for which PROTAGONISTS are responsible and which violate some NORM(S). The FE ACT identifies the act that is judged to be in or out of compliance with the norms. The FE NORM identifies the rules or norms that ought to guide a person's behavior. The FE PROTAGONIST refers to the person whose behavior is in or out of compliance with norms. Finally, the FE STATE_OF_AFFAIRS refers to the situation that may violate a law or rule (see Boas 2005a).

Applying the principles of Frame Semantics to the description and analysis of the English lexicon, the FrameNet project (Lowe et al. 1997, Baker et al. 1998) at the International Computer Science Institute in Berkeley, California, is in the process of creating a database of lexical entries for several thousand words taken from a variety of semantic domains. Based on data from the British National Corpus and other corpora, FrameNet identifies and describes semantic frames and analyzes the meanings of words by appealing directly to the frames that underlie their meaning. In addition, it studies the syntactic properties of words by asking how their semantic properties are given syntactic form (Fillmore et al. 2003a: 235). Between 1997 and 2008, FrameNet defined close to 7,000 lexical units (LUs) (a word in one of its senses) in more than 900 frames.

The workflow of FrameNet begins by defining frame descriptions (based on corpus evidence) for the words to be analyzed. Then, the following steps are taken: "(1) characterizing schematically the kind of entity or situation represented by the frame, (2) choosing mnemonics for labeling the entities or components of the frame, and (3) constructing a working list of words that appear to belong to the frame, where membership in the same frame will mean that the phrases that contain the LUs will all permit comparable semantic analyses" (Fillmore et al. 2003b: 297). The next step focuses on finding corpus sentences in the British National Corpus that illustrate typical uses of the target words in specific frames. Then, these corpus sentences are extracted mechanically and annotated manually by tagging the FEs realized in them. At last, lexical entries are automatically prepared and stored in the database (for more details, see Fillmore and Atkins 1998 and Fillmore 2003b).

Users accessing the FrameNet data on-line may use different types of search interfaces that allow searches by lexical unit (LU) or by semantic frames.[12] Lexical entries in FrameNet are structured as follows: They offer

12. This section is based on Boas (2005a). The FrameNet data can be accessed online at [http://framenet.icsi.berkeley.edu].

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 16)

a link to the definition of the frame to which the LU belongs, including FE definitions, and example sentences exemplifying prototypical instances of FEs. In addition, the FrameNet database includes a list of all LUs that evoke the frame, and provides for each frame-specific information about various frame-to-frame relations (e.g., child-parent relation and sub-frame relation (see Fillmore et al. 2003b)).

The central component of a lexical entry of a LU in FrameNet consists of three parts. The first provides the Frame Element Table (a list of all FEs found within the frame) and corresponding annotated corpus sentences demonstrating how FEs are realized syntactically. Note that FrameNet uses different colors to highlight each FE, making it easier to identify individual FEs. Due to formatting restrictions, FE names are not color-coded in Figures 7–9.

Figure 7 illustrates how FEs in the FE table and the corresponding annotated corpus sentences are displayed for the LU *comply*. In this part, words or phrases instantiating certain FEs in the annotated corpus sentences are annotated with the same FE name as in the FE table above them. This type of display allows users to identify the variety of different FE instantiations across a broad spectrum of words and phrases.

Notice the split of annotated corpus sentences into different groups according to different types of combinations of FEs. Numbers in the table represent the total number of annotated example sentences in FrameNet. Numbers at the beginning of each annotated example sentence represent their location in the British National Corpus. For example, in the first annotated example sentence in Figure 7 *comply*, which is the target ("Tgt") evoking the `Compliance` frame, occurs with the FEs Act, Degree, and Norm, while in the second example sentence it occurs only with Act and Norm. The numbers at the beginning of sentences show where each sentence occurs in the British National Corpus. FE names are displayed in terms of subscript notations following the first square bracket.

Next, consider Figure 8, which illustrates the second part of a lexical entry in FrameNet, namely the Realization Table of the Lexical Entry Report. Besides providing a dictionary definition of the relevant LU, in this case *comply*, it summarizes the different syntactic realizations of the frame elements. In the left column we find the names of different core FEs (Act, Norm, Protagonist, and State_of_Affairs), in the middle column we see the number of annotated example sentences in FrameNet, and in the right column we find the different types of syntactic realizations of the respective FEs. Consider the FE Norm, which appears 23 times, 21 of those times as a prepositional phrase headed by *with*, once as a definite null in-

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 17)

Num	**FE/LUset** (sort = FE; Compliance, comply, V,)
01	Act + Degree + **comply.V** + Norm
02	Act + **comply.V** + Norm
01	Norm + **comply.V** + (Protagonist)
03	Protagonist + **comply.V** + Degree + Norm
01	Protagonist + **comply.V** + Manner + Norm
10	Protagonist + **comply.V** + Norm
01	Protagonist + **comply.V** + Norm + Time
01	State_of_Affairs + **comply.V** + Norm
01	State_of_Affairs + **comply.V** + (Norm)
02	**comply.V** + Norm + (Protagonist)
23	

01. : Act + Degree + **comply.V** + Norm

1. 123614: [<Act> The last minute addition of the recommendation] did not [<Degree> in any way] *comply*^{Tgt} [<Norm> with the law] and the recommendation would be quashed.

02. : Act + **comply.V** + Norm

1. 123626: The court was told that [<Act> her appearance before the registrar] was solely to *comply*^{Tgt} [<Norm> with the formalities of Scots law].
2. 123758: [<Act> Spending by public sector organisations] has to *comply*^{Tgt} [<Norm> with complex and changing legal regulations], and is exposed to scrutiny at a number of levels.

01. : Norm + **comply.V** + (Protagonist)

1. 123932: If [<Norm> this rule] is not *complied*^{Tgt} [<Norm> with], the issuer is guilty of an offence, any subsequent contract etc entered into may be unenforceable and the issuer of the advertisement may face criminal charges and/or fines. [<Protagonist> CNI]

Figure 7. First part of FrameNet entry for *comply*

Comply.v

Frame: Compliance

Definition: COD: act in accordance with a wish or command

The Frame elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realizations(s)
Act	(3)	NP.Ext (3)
Norm	(23)	PP[with].Dep (21) DNI.–(1) NP.Ext (1) PP[to].Dep (1)
Protagonist	(18)	CNI.–(3) NP.Ext (15)
State of Affairs	(2)	NP.Ext (2)

Figure 8. FrameNet entry for *comply*, Realization Table

stantiation (DNI), once as an external noun phrase argument, and once as a prepositional phrase headed by *to* (for details see Boas 2005b).

The third part of the Lexical Entry Report summarizes the valence patterns found with a LU, that is, "the various combinations of frame elements and their syntactic realizations which might be present in a given sentence" (Fillmore et al. 2003a: 330). The third column from the left in the valence table for *comply* in Figure 9 illustrates how the FE Norm may be realized in terms of two different types of external arguments: either as an external noun phrase argument, or as an external prepositional phrase headed by *with*. Clicking on the link (in this case "3" or "1") in the column to the left of the valence patterns leads the user to a display of annotated examples sentences illustrating the valence pattern (see Figure 7 above).[13]

13. FEs which are conceptually salient but do not occur as overt lexical or phrasal material are marked as null instantiations. There are three different types of null instantiation: Constructional Null Instantiation (CNI), Definite Null Instantiation (DNI), and Indefinite Null Instantiation (INI). See Fillmore et al. (2003b: 320–321) for more details.

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 19)

Valence Patterns

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns		
3 TOTAL	**Act**	**Norm**	
(3)	NP Ext	PP[with] Dep	
1 TOTAL	**Norm**	**Norm**	**Protagonist**
(1)	NP Ext	PP[with] Dep	CNI –
16 TOTAL	**Norm**	**Protagonist**	
(2)	PP[with] Dep	CNI –	
(14)	PP[with] Dep	NP Ext	
1 TOTAL	**Norm**	**Protagonist**	**Protagonist**
(1)	PP[with] Dep	NP Ext	NP Ext
2 TOTAL	**Norm**	**State_of_Affairs**	
(1)	DNI –	NP Ext	
(1)	PP[to] Dep	NP Ext	

Figure 9. Partial FrameNet entry for *comply*, Valence Table

FrameNet differs from other approaches to lexical description such as WordNet (Fellbaum 1998) in that it makes use of independent organizational units that are larger than words, i.e., semantic frames (see also Atkins 2002, Ohara et al. 2003, Boas 2005b, Atkins and Rundell 2008). As such, FrameNet facilitates a comparison of the comprehensive lexical descriptions and their manually annotated corpus-based example sentences with those of other LUs (also of other parts of speech) belonging to the same frame. Another advantage of the FrameNet architecture lies in the way lexical descriptions are related to each other. Using detailed semantic frames which capture the full background knowledge evoked by all LUs of the same frame makes it possible to systematically compare and contrast their numerous syntactic valence patterns (see Atkins 2002, Boas 2005a).

5. The structure and development of multilingual FrameNets

I now turn to an outline of the individual chapters in this volume. The main chapters provide a state-of-the-art implementation of the FrameNet methodology for the description and analysis of languages other than English. The FrameNets for other languages described in this volume vary from the original Berkeley FrameNet in the following points:

(1) Projects such as SALSA (see Burchardt et al., this volume) are interested in full-text annotation of an entire corpus instead of finding isolated corpus sentences to identify lexicographically relevant information as is the case with the Berkeley project, Spanish FrameNet (see Subirats, this volume), or the Romance FrameNet initiative;[14]

(2) FrameNets use different types of resources as data pools. That is, besides exploiting a mono-lingual corpus as is the case with Japanese FrameNet (see Ohara, this volume), projects such as French FrameNet (Pitel, this volume) also employ multi-lingual corpora and other existing lexical resources (see Fontenelle, this volume);

(3) FrameNets for other languages differ in the tools for corpus searches and annotation. While the Japanese and Spanish FrameNets choose to adopt the Berkeley FrameNet software (Baker et al. 2003) with slight modifications, others such as SALSA develop their own to conduct semi-automatic annotation on top of existing syntactic annotations found in the TIGER corpus, or they integrate off-the shelf software packages as is the case with French FrameNet or Hebrew FrameNet (Petruck, this volume);

(4) FrameNets focus on different semantic domains. While the majority of non-English FrameNets aim to create databases with broad coverage, other projects such as the *Kicktionary* (Schmidt, this volume) focus on specific lexical domains such as football language or terminology from bio-technology (see Dolbey et al. 2006);

(5) To produce parallel lexicon fragments for other languages, projects utilize different methodologies. While German FrameNet (Boas 2001, 2002) and Japanese FrameNet (Ohara, this volume) rely on manual

14. See http://www.icsi.berkeley.edu/~vincenzo/rfn/index.html.

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 21)

annotations, French FrameNet and BiFrameNet (Fung and Chen 2004) use semi-automatic and automatic approaches to create parallel lexicon fragments for French and Chinese.

To highlight the similarities and differences between the Berkeley FrameNet and other FrameNets, this volume is divided into four thematic sections. Chapters 1–3 offer an introduction to the basic concepts underlying the development of FrameNets for other languages, further expanding the initial proposals emerging from the DELIS project discussed in the previous section (Heid 1996a). Fontenelle's chapter *A bilingual lexical database for Frame Semantics* (a reprint of his 2000 *International Journal of Lexicography* paper) demonstrates how a FrameNet-type lexical database can be derived from an existing bilingual English-French dictionary. This contribution is significant, because it is the first to suggest (1) using the collocational information contained in the Collins-Robert bilingual machine readable dictionary to derive parallel lexicon fragments, and (2) combining Fillmore's Frame Semantics (Fillmore 1985) with Mel'čuk's lexical functions (Mel'čuk et al. 1988) in order to identify core frame elements, together with their syntax (see Alonso-Ramos 2003 and Bouveret and Fillmore 2008 for similar approaches). Fontenelle also shows how the database organization of the computational database makes it possible to readily access combinatorial information that is implicit and relevant to translation.

Boas' chapter *Semantic frames as interlingual representations for multilingual lexical databases* (a reprint of his 2005 *International Journal of Lexicography* paper) first discusses some of the key problems in the construction of multi-lingual lexical databases, such as polysemy, differences in syntactic and semantic valence patterns, differences in lexicalization patterns, and measuring paraphrase relations and translation equivalents. Based on the architecture of the English FrameNet database (Fillmore et al. 2003), it then suggests how FrameNet tools can be re-used to construct FrameNets for Spanish, German, and Japanese. Comparing some parallel Spanish lexicon fragments that result from this workflow, Boas' chapter demonstrates how parallel FrameNet entries differ from those of other multilingual lexical databases: (1) they provide for each entry an exhaustive account of the semantic and syntactic combinatorial possibilities of each lexical unit; (2) they offer for each entry semantically annotated example sentences from large electronic corpora, and (3) by employing semantic frames as interlingual representation, the parallel FrameNets make use of independently existing concepts that can be empirically verified.

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 22)

Schmidt's *The Kicktionary – a multilingual lexical resource of football language* directly implements the ideas proposed by Boas in the previous chapter. Schmidt describes the creation of an experimental tri-lingual FrameNet database (English-German-French) for a specific lexical domain, namely soccer (football) words. This FrameNet-type approach is different from other FrameNets in that it utilizes publicly available corpora from the world soccer organization (FIFA), which are available for a number of different languages. This contribution first shows how soccer texts in different languages are prepared for cross-linguistic comparison using a keyword-in-context program for parallel corpora. Then, it discusses how different lexicalization patterns found in the three languages influence the creation of parallel lexicon-fragments for soccer words, using FrameNet tools. Finally, this chapter addresses the question of polysemy and coverage of specific word senses (technical vocabulary) when dealing with domain-specific words in the creation of multi-lingual FrameNets.

Chapters 4–6 describe the different methods used for creating broad-coverage FrameNets for typologically diverse languages. While the Spanish, Japanese, and Hebrew FrameNet projects adopted the design and workflow of the original Berkeley FrameNet, they each differ with respect to the types of resources and tools used. They also vary in that each project has to address language-specific issues such as lexicalization patterns or frame composition. The discussion of a variety of language-specific phenomena demonstrates that it is not always possible to straightforwardly create parallel lexicon fragments on the basis of English FrameNet frames and lexical entries alone.

Subirats' chapter *Spanish FrameNet: A frame semantic analysis of the Spanish lexicon* demonstrates the re-usability of the English FrameNet tools for the creation of a lexical database for Spanish verbs, nouns, and adjectives. It first discusses the compilation of a 300-million word corpus (including both New World and European Spanish texts) for annotation purposes and the tagging of the corpus. It then describes the output of a tagger, which is a set of deterministic automata, one per corpus sentence, whose transitions are tagged with the lexical and morphological information of the word form in the electronic dictionary. Finally, it explains the extraction and subcorpora creation processes which provide annotators with examples of each possible syntactic configuration in which a lexical item can occur. Part two of Subirats' chapter shows how the English-based FrameNet tools (annotation software and database structure) are re-used for the creation of Spanish lexical entries, and how parallel lexical entries can be linked to each other. Finally, part three analyzes differences

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 23)

in lexicalization patterns in the communication and motion domains in order to show how such linguistic differences influence the design of the Spanish FrameNet database.

Ohara's *Frame-based contrastive lexical semantics in Japanese FrameNet: The case of 'risk' and 'kakeru'* explains the tools, resources, and workflow of the Japanese FrameNet project, which aims at creating a Japanese lexicon based on Frame Semantics. It first discusses in detail a number of technical issues that arise when re-using English FrameNet tools for the description of a non-Indo-European language: compilation of a Japanese corpus suitable for annotation purposes, assignment of morphological and sentence boundaries, and development of an annotation tool for Japanese. Then, the chapter addresses some of the linguistic problems with applying frame-semantic categories to the description of Japanese: (1) how to identify and capture multiple senses and uses associated with a single form, (2) how to deal with recognized differences in senses and conditions of use among verbs related in meaning, and (3) how to create Japanese-specific frames for cases in which English-based frames are not fine-grained enough to capture some of the relevant semantic distinctions made in Japanese. Finally, the paper shows how Japanese lexicon fragments can be systematically linked to their English counterparts.

Petruck's chapter *Typological considerations in constructing a Hebrew FrameNet* illustrates the challenges faced when creating a FrameNet resource for a Semitic language. It first discusses how Hebrew FrameNet is aimed at documenting the range of semantic and syntactic combinatorial possibilities (*valences*) of each word in each of its senses by annotating example sentences and compiling the results for display. It then examines how full-text annotation of frame evoking elements (FEEs) for an existing newspaper corpus are created in order (1) to develop the infrastructure for using the FrameNet Desktop for the analysis of Hebrew texts and (2) to investigate at what level of linguistic description and computational representation the lexicon of contemporary Hebrew can be characterized in the same terms as the lexicon of English, thereby necessarily considering the matter of transferability of FrameNet machinery to a language other than English. The investigation of how events and scenarios are expressed through the same or different frames illustrate the different lexicalization patterns of Hebrew and English (Talmy 2000), thus contributing to cross-linguistic studies as well.

Chapters 7–8 address the question of how parts of the FrameNet workflow can be automated when creating FrameNets for other languages. This is an important issue because the current workflow of the Berkeley

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 24)

project is time and labor intensive due to its reliance on the manual creation of frames as well as the manual annotation of corpus examples.[15] The chapter *Using FrameNet for the semantic analysis of German: annotation, representation, and automation* by Burchardt et al. discusses the tools, workflow, annotation practices, and goals of the *Saarbrücken Lexical Semantics Acquisition* (SALSA) Project, which creates a FrameNet-type lexical database for German. One of the significant outcomes of SALSA is that the English frames and FEs developed by the Berkeley project for English can be re-used fortuitously to describe German predicate-argument structures. SALSA differs from the English FrameNet design and workflow in that it annotates all frame-evoking words in an entire corpus (the German TIGER corpus) thereby maximizing both annotation consistency and coverage. This is in contrast to the Berkeley FrameNet, which focuses on lexicographically relevant examples from the BNC. The chapter details the treatment and annotation of limited compositionality phenomena such as support verb constructions, idioms, and metaphors. This chapter also demonstrates how SALSA investigates several options for acquiring a semantic lexicon semi-automatically, including shallow semantic parsing. Finally, this chapter addresses some typological differences (vagueness, ambiguity, verb class membership, cross-linguistic paraphrase modeling, etc.) that arise when applying English-based semantic frames to the description of German words.

Pitel's chapter on *Cross-lingual labeling of semantic predicates and roles: A low-resource method based on bilingual l(atent) s(emantic) a(nalysis)* examines how existing FrameNet tools (annotation software and database) can be adapted for the creation of a French FrameNet. Besides discussing linguistic-typological and technical issues that arise during this process, this chapter focuses on the question of how the modified tools and resulting lexical entries for French can be re-used for other Romance languages such as Italian, Romanian, Portuguese, and Catalan, which are currently being analyzed by the Romance FrameNet consortium (inspired by MultiSemCor). The goal of this effort is to (1) create a consistent aligned and frame-annotated multi-lingual corpus; (2) highlight cross-language regularities, and structural intra- and extra-typological idiosyncrasies; (3) create a semantically indexed translation memory and an inverse multi-lingual dictionary; (4) create one of the first freely available resources that contains cross-

15. Note that some proposals have been put forward for automatically inducing frame semantic verb classes in English (see Green and Dorr 2004, Green et al. 2004).

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 25)

languages sub-categorization and collocational mappings; (5) reuse the work done on automatic role assignment and semantic parsing.

The last two chapters offer different perspectives on multilingual computational lexicography that go beyond the methodology underlying the various FrameNet-like projects. Farwell et al.'s *Interlingual annotation of multilingual text corpora and FrameNet* offers a fresh look at the usability of multilingual annotated corpora for inducing FrameNet-type lexicon fragments for a variety of languages. The chapter describes the annotation process being used in a multi-site project to create six sizable bilingual parallel corpora annotated with a consistent interlingua representation. The authors examine the multilingual corpora (as well as the three stages of interlingual representation being developed), the annotation process, and the methodology for evaluation the interlingual representations. The resulting interlingual representations are then compared with the semantic frames and lexical entries of the FrameNet database in order to discuss the differences and their implications for natural language processing tasks, such as machine translation, question answering, and information extraction.

The final chapter *Universals and idiosyncrasies in multilingual WordNets* by Vossen and Fellbaum addresses design issues surrounding the use of an interlingual index for mapping between lexical databases for different languages as opposed to semantic frames. Building on prior results, the authors propose an extension of the EuroWordNet model (Vossen 1998) to cover a large number of languages (including lesser-known ones), in the "Global WordNet Grid" (GWG). Vossen and Fellbaum envision that the GWG will include an ontology as the basis for a universal concept index and that it will allow the large-scale empirical investigation of fundamental theoretical questions. This enterprise will eventually reveal which lexicalizations are universal or idiosyncratic and how they can be linked to the universal concept index. Finally, the authors offer a comparison of the linguistic-typological differences between multilingual WordNets and multilingual FrameNets, thereby highlighting the different goals of the two approaches.

References

Alberto, P. and P. Bennett (eds.)
1995 *Lexical issues in machine translation.* Studies in Machine Translation and Natural Language Processing, Vol. 8. Luxembourg: European Commission.

Alonso-Ramos, M.
2003 Éléments du frame vs. Actants de l'unité lexicale. In: *MTT 2003 – Proceedings of the First International Conference on Meaning-Text Theory*, 77–88. Paris: École Normale Supérieure.

Altenberg, B. and S. Granger (eds.)
2002 *Lexis in contrast.* Amsterdam/Philadelphia: John Benjamins.

Amsler, R.A.
1980 *The structure of the Merriam-Webster Pocket Dictionary*. Ph.D. dissertation, The University of Texas at Austin.

Antoni-Lay, M.-H., G. Francopoulo and L. Zaysser
1994 A generic model for reusable lexicons: The GENELEX project. *Literary and Linguistic Computing* 9(1), 47–54.

Atkins, B.T.S.
1993 The contribution of lexicography. In: Bates, M. and R.M. Weischedel (eds.), *Challenges in Natural Language Processing*, 37–75. Cambridge: Cambridge University Press.

Atkins, B.T.S.
2002 Then and now: competence and performance in 35 years of lexicography. In: *EURALEX 2002 Proceedings.* Reprinted in Fontenelle, T. (ed.), *Practical Lexicography – A Reader*. Oxford: Oxford University Press (2008).

Atkins, B.T.S. and A. Duval
1978 *Robert and Collins Dictionnaire Français-Anglais, Anglais-Français.* Paris: Le Robert/Glasgow: Collins.

Atkins, B.T.S., J. Kegl and B. Levin
1986 Explicit and implicit information in dictionaries. In: *Lexicon Project Working Papers 12*, Center for Cognitive Science, MIT, Cambridge, MA.

Atkins, B.T.S. and B. Levin
1991 Admitting impediments. In: U. Zernik, (ed.), *Lexical Acquisition Using Online Resources to Build a Lexicon*, 233–262. Hillsdale: Lawrence Erlbaum Associates.

Atkins, B.T.S and M. Rundell
2008 *Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Atkins, B.T.S. and A. Zampolli (eds.)
1994 *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Baker, C.F., C.J. Fillmore and J.B. Lowe
1998 The Berkeley FrameNet Project. In: *COLING-ACL '98: Proceedings of the Conference*, 86–90.

Baker, C.F., C.J. Fillmore and B. Cronin
2003 The structure of the FrameNet database. *International Journal of Lexicography* 16, 281–296.

Béjoint, Henri
1994 *Tradition and Innovation in Modern English Dictionaries*. Oxford: Clarendon Press.

Béjoint, Henri
2001 *Modern Lexicography*. Oxford: Oxford University Press.

Bennet, W.S. and J. Slocum
1985 The LRC machine translation system. *Computational Linguistics* 11(2–3), 111–121.

Benson, P.
2001 *Ethnocentrism and the English Dictionary*. London: Routledge.

Boas, Hans C.
2001 Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography. In: P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.), *Proceedings of Corpus Linguistics 2001*, 64–73.

Boas, Hans C.
2002 Bilingual FrameNet dictionaries for machine translation. In: M. González Rodríguez and C. Paz Suárez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation*, Vol. IV, 1364–1371. Las Palmas, Spain.

Boas, Hans C.
2005a Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography* 18(4), 445–478.

Boas, Hans C.
2005b From theory to practice: Frame Semantics and the design of FrameNet. In: S. Langer and D. Schnorbusch (eds.), *Semantik im Lexikon*, 129–160. Tübingen: Narr.

Boguraev, B. and T. Briscoe
1989 *Computational Lexicography for Natural Language Processing*. London and New York: Longman.

Bouveret, M. and C.J. Fillmore
2008 Matching verbo-nominal constructions in FrameNet with lexical functions in MTT. In: E. Bernal and J. De Cesaris (eds.) *Euralex 2008 Proceedings*, 297–308. Barcelona.

Calzolari, N.
1991 Lexical databases and textual corpora: perspectives of integration of a lexical knowledge base. In: U. Zernik (ed.), *Lexical acquisition: exploiting on-line resources to build a lexicon*, 191–208. Hillsdale: Lawrence Erlbaum.

Calzolari, N. and T. Briscoe
1995 ACQUILEX-I and –II: Acquisition of lexical knowledge from machine readable dictionaries and text corpora. *Cahiers Lexicologique* 67(2), 95–114.

Calzolari, N., R, Grishman, M. Palmer, B.T.S. Atkins, N. Bel, F. Bertagna, P. Bouillon, B. Dorr, C. Fellbaum, D. Gibbon, N. Habash, E. Lange, S. Lehmann, A. Lenci, S. McCormick, J. McNaught, A. Ogonowski, J. Pentheroudakis, S. Richardson, G. Thurmair, L. Vanderwende, M. Villegas, P. Vossen and A. Zampolli.
2001 *Survey of major approaches towards bilingual/multilingual lexicons*. ISLE Computational Lexicons Working Group Deliverable D2.1–D3.1. Online: http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm.

Calzolari, N., F. Bertagna, A. Lenci and M. Monachini, with S. Atkins, N. Bel, - P. Bouillon, T. Charoenporn, D. Gibbon, R. Grishman, C.-R. Huang, A. Kawtrakul, N. Ide, H-Y.Lee, P.J.K. Li, J. McNaught, J. Odijk, M. Palmer, V. Quochi, R. Reeves, D.M. Sharma, V. Sornlertlamvanich, T. Tokunaga, G. Thurmair, M. Villegas, A. Zampolli and El Zeiton.
2003 *Standards and best practice for multilingual computational lexicons and MILE (the multilingual ISLE lexical entry)*. Deliverable D2.2–D3.2, ISLE Computational Lexicon Working Group. Online at http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm.

Copestake, A.
1992 *The Representation of Lexical Semantic Information*. Ph.D. dissertation, University of Sussex.

Copestake, A. and A. Sanfilippo
1993 Multilingual Lexical Representation. Paper presented at the *AAAI Spring Colloquium on Building Lexicons for Machine Translation*. Stanford, CA. *ACQUILEX II Working Papers No. 3*.

Cruse, A.
1986 *Lexical Semantics*. Cambridge: Cambridge University Press.

Dolbey, A., M. Ellsworth, and J. Scheffczyk
2006 BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. Paper presented at the *International Workshop Biomedical Ontology in Action*, November 8, 2006, Baltimore, MD.

Durand, J., P. Bennett, V. Allegranza, F. Van Eynde, L. Humphreys, P. Schmidt, and E. Steiner
1991 The Eurotra Linguistic Specifications: an overview. In: *Machine Translation 6*, 103–147. Dordrecht: Kluwer.

Emele, M.
1993 TFS – The typed feature structure representation formalism. In: H. Uszkoreit (ed.), *Proceedings of the EAGLES workshop on implemented formalisms*. Saarbrücken: DFKI-Report.

Emele, M. and U. Heid
1994 *Delis*: tools for corpus based lexicon building. In: *Proceedings of Konvens-94*, (Heidelberg: Springer) 1994, [=Informatik Xpress 6].

Fellbaum, C.
1998 *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, C.J.
1982 Frame Semantics. In: Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*, 111–138. Seoul: Hanshin.

Fillmore, C.J.
1985 Frames and the Semantics of Understanding. *Quadernie di Semantica* 6(2), 222–254.

Fillmore, C.J. and B.T.S. Atkins
1992 Towards a frame-based lexicon: The semantics of RISK and its neighbors. In: A. Lehrer and E. Kittay (eds.), *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, 75–102. Hillsdale: Erlbaum,

Fillmore, C.J. and B.T.S. Atkins
1994 Starting where the dictionaries stop: The challenge for computational lexicography. In: B.T.S. Atkins. and A. Zampolli (eds.), *Computational Approaches to the Lexicon*, 349–393. Oxford: Oxford University Press.

Fillmore, C.J. and B.T.S. Atkins
1998 FrameNet and lexicographic relevance. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain.

Fillmore, C.J. and B.T.S. Atkins
2000 Describing polysemy: The case of *crawl*. In: Y. Ravin and C. Leacock (eds.), *Polysemy*, 91–110. Oxford: Oxford University Press.

Fillmore, C.J. and M. Petruck
2003 FrameNet Glossary. *International Journal of Lexicography* 16(3), 359–361.

Fillmore, C.J., C.R. Johnson and M. Petruck
2003a Background to FrameNet. *International Journal of Lexicography* 16(3), 235–250.

Fillmore, C.J., M. Petruck, J. Ruppenhofer and A. Wright
2003b FrameNet in action: The case of attaching. *International Journal of Lexicography* 16(3), 297–332.

Fontenelle, T.
1997 *Turning a Bilingual Dictionary into a Lexical Semantic Database*. Tübingen: Niemeyer.

Fontenelle, T.
2008 Linguistic research and learners' dictionaries: the Longman Dictionary of Contemporary English. In: A.P. Cowie (ed.), *Oxford History of English Lexicography*, 412–435. Oxford: Oxford University Press.

Fung, P. and B. Chen
2004 BiFrameNet: Bilingual frame semantics resource construction by cross-lingual induction. In *Proceedings of COLING 2004*. Geneva, Switzerland.

Gerber, L. and J. Young
1997 SYSTRAN MT Dictionary Development. Paper presented at the *MT Summit*, San Diego.

Green, J.
1996 *Chasing the Sun: Dictionary-makers and the Dictionaries they made*. London: Pimlico.

Green, R. and B. Dorr
2004 Inducing a Semantic Frame Lexicon from WordNet Data. In: *Proceedings of the Workshop on Text Meaning and Interpretation, Association for Computational Linguistics*, Barcelona, Spain, 2004.

Green, R., B. Dorr and P. Resnik
2004 Inducing Frame Semantic Verb Classes from WordNet and LDOCE. In: *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*.

Hamp, B. and H. Feldweg
1997 GermaNet: a lexical-semantic net for German. In: P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo and Y. Wilks (eds.), *Proceedings of the ACL/EACL-97 Workshop on automatic information extraction and building of lexical semantic resources for NLP applications,* Madrid, 9–15.

Hartmann, R.R.K. and G. James
1998 *Dictionary of Lexicography*. London/New York: Routledge.

Heid, U.
1996a On the verification of lexical descriptions in text corpora. In: N. Weber (ed.): *Semantik, Lexikographie und Computeranwendungen*, 289–306. Tübingen: Niemeyer.

Heid, U.
1996b Creating Multilingual Data Collection for Bilingual Lexicography from Parallel Monolingual Lexicons. In: *Proceedings of Euralex 1996*, Göteburg University.

Heid, U.
1997 *Zur Strukturierung von einsprachigen und mehrsprachigen kontrastiven elektronischen Wörterbüchern*. Tübingen: Niemeyer.

Heid, U.
2006 Valenzwörterbücher im Netz. In: P. Steiner, H.C. Boas and S. Schierholz (eds.), *Contrastive Studies and Valency*, 69–89. *Studies in Honor of Hans Ulrich Boas.* Frankfurt/New York: Peter Lang.

Heid, U., W. Martin and I. Posch
1991 *Feasibility and standards for the collocational description of lexi-*

cal items. Stuttgart and Amsterdam, EUROTRA-7 Study, Document DOC-9/4.

Heid, U. and J. McNaught
1991 *EUROTRA – Feasibility and Project Definition Study on the Reusability of lexical and terminological resources in Computerized Applications – Final Report* Stuttgart/Luxembourg: IMS-CL/ Kommission der europäischen Gemeinschaften.

Johnson, R., M. King and L. des Tombe
1985 EUROTRA: A multilingual system under development. *Computational Linguistics* 11(2–3): 155–169.

Johnson, R., M. King and L. des Tombe
2003 EUROTRA: Computational techniques. In: S. Nirenburg, H. Somers, and Y. Wilks (eds.), *Readings in Machine Translation*, 345–350. Cambridge, MA: MIT Press.

Kunze, C. and L. Lemnitzer
2002 GermaNet – representation, visualization, application. In: *LREC 2002 Proceedings* Vol. V.: 1465–1491.

Landau, S.I.
1989 *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Lehmann, W.P.
1998 Machine Translation at Texas: The Early Years. Online at http://www.utexas.edu/cola/centers/lrc/mt/earlymt.html.

Lowe, J.B., C.F. Baker and C.J. Fillmore
1997 A frame-semantic approach to semantic annotation. In: *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* held April 4–5, in Washington, D.C., USA in conjunction with ANLP–97.

Makkai, A.
1980 Theoretical and Practical Aspects of an Associative Lexicon for 20th Century English. In: L. Zgusta, L. (ed.), *Theory and Method in Lexicography: Western and Non-Western Perspectives*, 125–46. Columbia, SC: Hornbeam Press.

McNaught, J.
1988 Computational Lexicography and Computational Linguistics. *Lexicographica* 4, 19–33.

Mel'čuk, I., N. Arbatchewsky-Jumarie, L. Dagenais, L. Elnitsky, L. Iordanskaja, M.-N. Lefebvre and S. Mantha.
1988 *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches Lexico-sémantiques*. Montréal: Les Presses de l'Université de Montréal.

Michiels, A.
1982 *Exploiting a Large Dictionary Database*. Ph.D. dissertation, University of Liège.

Miller, G., et al.
1990 Five Papers about WordNet. In: *CSL-Report* 43, Cognitive Science Laboratory, Princeton University.

MULTILEX (ed.)
1993 *Standards for a Multifunctional Lexicon*, CAP GEMINI INNOVATION for the MULTILEX Consortium, Paris.

Ooi, Vincent
1998 *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Papegaaij, B.C., V. Sadler and A.P.M. Witkam (eds.)
1986 *Word Expert Semantics: An Interlingual Knowledge-based Approach*. Dordrecht: Foris.

Peters, W., I. Peters and P. Vossen
1998 The reduction of semantic ambiguity in linguistic resources. In: A. Rubio, N. Gallardo, R. Catro, and A. Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*, 409–416. Granada.

Petruck, M.R.L.
1996 Frame Semantics. In: J. Verschueren, J.-O. Östman, J. Blommaert and C. Bulcaen (eds.), *Handbook of Pragmatics*, 1–13. Amsterdam/Philadelphia: Benjamins.

Pollard, C. and I. Sag
1994 *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Procter, P. (ed.)
1978 *Longman Dictionary of Contemporary English* (1st edition). Harlow: Longman.

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

Ohara, K., S.K. Fujii, H. Saito, S. Ishizaki, T. Ohori and R. Suzuki
2003 The Japanese FrameNet Project: A preliminary report. In: *Proceedings of the Pacific Association for Computational Linguistics* (PACLING03), 249–254.

Ramsay, A.M.
1991 Artificial Intelligence. In: K. Malmkjær (ed.), *The Linguistics Encyclopedia*, 28–38. London: Routledge.

Slocum, J.
2006 *Machine translation at Texas: The later years*. Online at http://www.utexas.edu/cola/centers/lrc/mt/latermt.html.

Summers, D.
1987 *Longman Dictionary of Contemporary English* (2nd edition). Harlow: Longman.

Svensen, B.
1993 *Practical Lexicography*. Oxford: Oxford University Press.

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 33)

Talmy, L.
2000 *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Vossen, P.
1997 EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, March 5–7, 1997, Zurich.

Vossen, P.
1998 (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks for European Languages*. Dordrecht: Kluwer.

Vossen, P.
2001 Condensed meaning in EuroWordNet. In: P. Bouillon and F. Busa (eds.), *The language of word meaning*, 363–383. Cambridge: Cambridge University Press.

Vossen, P.
2004 EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Lexicography* 17(2), 161–173.

Vossen, P., W. Peters and P. Díez-Orzaz
1997 The Multilingual design of the EuroWordNet Database. In: K. Mahesh (ed.), *Ontologies and multilingual NLP, Proceedings of workshop at IJCAI-97*, Nagoya, Japan, August 23–29.

Walker, D., A. Zampolli and N. Calzolari (eds.)
1995 *Automating the Lexicon: Research and Practice in a Multilingual Environment*. Oxford: Oxford University Press.

Zampolli, A.
1991 Technology and linguistic resources. In: M. Katzen (ed.), *Scholarship and Technology in the Humanities*. London: British Library Research.

Zampolli, A.
1994 Introduction. In: B.T.S. Atkins and Z. Zampolli (eds.), *Computational Approaches to the Lexicon*, 3–16. Oxford: Oxford University Press.

Zgusta, L.
1971 *Manual of Lexicography*. The Hague: Mouton.

(V9 20/5/09 02:48) WDG (155mm×230mm) TimesNRMT 1114 Boas (AC1) pp. 1–34 1114 Boas_01_Ch01 (p. 34)