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). One of the results of computational lexicography is a dramatic enhancement of Natural Language Processing (NLP) systems through richer machine-readable dictionaries (Boguraev and Bricc1ce 1989). One early example is the machine-readable version of the Longman Dictionary of Contemporary English (benchmarks: LDOCE; Procter 1978), which turned out to be particularly useful for NLP research because it offered detailed subcategorizations of major word classes (see Amater 1980, Michels 1982, Ovel 1998, and Fontenelle 2000).

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 (1923: 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, Kgl, and Levin 1986), unsystematic compiling of one single dictionary, ambiguities within entries, and incompatible compiling across dictionaries (Ahons and Levin 1991). Such problems — as well as new insights — lead lexicographers to revise and restructure MRDs, as, for example, has been done for...

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 Daval 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, Fennerelle 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 Papagou et al. 1986, Ooi 1998). However, this approach proved to be unsuccessful because the translation of words...
Recent trends in multilingual computational lexicography

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

6. These early exercises in developing MRD's for MT demonstrated the
7. prevalence of the "lexical acquisition bottleneck." To develop large-scale
8. lexical resources for multilingual NLP applications, there were in principle
9. two different approaches: (1) re-using existing resources, or (2) building
10. MRDs from scratch with the help of teams of trained lexicographers.
11. Over the next decades, several efforts were aimed at creating more so-
12. phisticated MRDs using these two methodologies. In what follows, I
13. present a brief overview of a select number of these efforts to set up
14. the context for our discussion of the design of multi-lingual FrameNets
15. in sections 4-5.
16. During the 1990s and 1960s, MRDs became more structured, partially
17. due to the development of more sophisticated syntactic parsing techniques
18. and the newly emerging designs of MT systems that made principled dis-
19. tinctions between linguistic rules, the grammar, and the lexicon (Lehmann
20. 1994). One system that employed such a design was the METAL transla-
21. tion system developed by the Linguistics Research Center at the Univer-
22. sity of Texas at Austin beginning in the 1960s, whose development contin-
23. ued (with various modifications) until the 1990s (see Sticlim 2000). To
24. produce German-to-English translations, the system relied on monoling-
25. ual dictionaries for English and German that were largely created from
26. scratch, each containing about 10,000 entries. The entries in the METAL
27. dictionary were indexed by canonical form (the usual spelling one finds in
28. a printed dictionary) (Bennett and Sticlim 1985). For the input of lexical
29. entries, a lexical default program was developed that allowed the lexic-
30. ographers to specify only minimal information about a particular entry such
31. as root form and lexical category. The program then heuristically encoded
32. most of the remaining necessary features and values. The METAL lexicon
33. included detailed morpho-syntactic information about part of speech, in-
34. factional class, gender, number, stem vs. count noun, and gradation.
35. With respect to syntax, the lexicon specified the subcategorization frame
36. and the types of auxiliaries. On the semantic side, the METAL lexicon
37. provided only minimal information, namely about the semantic type and
38. the domain (Calzolari et al. 2001: 108-109). The resulting MRD was
39. somewhat limited in scope – it was originally developed for technical
40. translations from German to English – but its minimal entry structure
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 some 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.

\section{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 Boisot 1995), which extracted lexical information from multiple MRDs in one 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
Recent trends in multilingual computational lexicography

sense the lexical knowledge base (LKB) contained phonological, morpho-
logical, syntactic, and semantic/pragmatic information capable of deploy-
ment in the lexical components of a wide variety of practical NLP sys-
tems. Figure 1 illustrates the structure of an entry in the LKB.

Figure 1. The LKB entry for chocolate (Copestake 1993)

Figure 1 shows that more detailed semantic information played an im-
portant role in ACQUILEX. Fustegoviy'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 Fentonelle 1997: 13).²
 Another project funded by the European Commission was EUROTRA-
7 (Heil 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 Gro-
seto Workshop), which showed that there was a growing need for standard-
dized and reusable lexical descriptions that could be employed independ-
ently 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 de-
tailed list of diverging and converging needs, which, led to a methodologi-
cal 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
facs, they made different generalizations using varying descriptive devices
(see Heil et al. 1991).
 To provide the various frameworks with reusable lexical and termino-
 logical 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 differ-
ence. 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 these main areas:
acquisition, representation, and application. The recommendations pro-

² For details on the LKB, see Copestake (1992) and Copestake and Santillipo
(1993).
³ Other projects building on the recommendations of EUROTRA-7 were
MULTILEX (MULTILEX 1993), and GENLEX (Antoni-Lay et al. 1994).
duced by EUROTRA-7 were significant for the development of future
multilingual lexical resources because they explicitly described (1) the ini-
tial specifications needed for a model of a robust lexicon, and (2) the
need for standardized formats allowing researchers from academia and in-
dustry to use the same lexical resources for a variety of applications, re-
gardless of their theoretical backgrounds.1

One of the follow-on 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 con-
cerned with standardization of multilingual computational lexicons, but
also grammar formalisms, evaluation and assessment, and spoken lan-
guage. The EAGLES working group on computational lexicons resulted
in a series of recommendations for devising standardized architectures for
multilingual lexicons.2 These recommendations were instrumental in the
design of the PAROLE-SIMPLE lexicons for twelve European languages
(Caltolari et al. 2001: 83), including the semantic lexicons with about
10,000 word meanings. To capture the various dimensions of word mean-
ing, the semantic representation relied on an extension of Postčevićky’s
(i953) ‘quasxic 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 (Caltolari et al. 2001: 83). The SIMPLE model in-
cluded three types of formal entries, 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
sense/s 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 (Caltolari et al. 2001: 83). The second formal en-
tity in the SIMPLE model was the (Semantic) Type, which represented the
semantic type assigned to SemUs. The four semantic types were organized

5. See http://www.itc-cri.it/EAGLES98/bronze.html and http://www.itc-
cri.de/EAGLES98/EAGLES9LE.PDF for details on the recommendations cre-
ated by EAGLES.
in terms of Pustejovsky's (1995) qualia structures, which in turn were categorized 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, predicate, 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
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 entity 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 "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
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 FrameNet 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-Linguat-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 synsets (the synsets, see Miller et al., 1990) such as hyponymy, antonymy, meronymy, etc. (see Cruse 1996). 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). 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 ItaWordNet, is linked to an unstructured ILI and a top concept ontology.

The ILI provides mapping across individual language WordNet structures and consists of a concrete/universal index of meaning (1044 fundamental concepts) (Vossen 2001, 2004). 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_SYNONYM hold.

6. In EuroWordNet, there are no concepts for which there are no words or expressions in a language. In contrast, GermaNet (Hartn and Feldweg 1997, Krute and Lemminger 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 as explained in Vossen et al. (1997: 1) is 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 metalexicon, capable of supplying a satisfactory conceptual backbone to all the languages.”
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 (which is linked to a top-level ontology) 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_SYNGNYM. 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)).
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 GermanNet for German (Kurze and Lemnitzer 2002) and a number of projects under the auspices of the Global WordNet Association. The current move towards a Global WordNet Grid (GWG) (Vossen and Fellbaum, this volume) seeks 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). 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 hierarchial modules for each of the descriptive levels encoded, i.e. Morphosyntax, Syntax, and Semantics (see Heid 1996a: 286).

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

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.
Recent trends in multilingual computational lexicography

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

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

<table>
<thead>
<tr>
<th>Construct →</th>
<th>Descriptive Devices</th>
<th>Constellations (Classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>ROLES</td>
<td>ROLE CONSTELLATIONS</td>
</tr>
<tr>
<td>Lexical syntax</td>
<td>GRAMM. FUNCTIONS</td>
<td>TOPMOST SYNTACTIC CLASSES</td>
</tr>
<tr>
<td>Functional syntax</td>
<td>SYNTACTIC CATEGORIES, PHRASE TYPES</td>
<td>SPECIFIC SYNTACTIC CLASSES</td>
</tr>
</tbody>
</table>

The second defining characteristic was that DELIS dictionaries were
classificatory in that the description of each level was organized in mono-
tonic 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 combina-
tions of semantic role, 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
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
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 back-
grounds of knowledge (semantic "frames") against which the meanings of words are interpreted. A "frame is a cognitive structuring device, parts of which are indexed by words associated with it and used in the service of understanding" (Petrucci 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 con-
ceptual 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 stems 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 
comply, subdue, adhere to, comply, com-
pliance, and violate, among others (Johnson et al. 2003). The Compliance frame represents a kind of situation in which different types of relation-
ship hold between "Frame Elements" (FEs), which are defined as situ-
tion-specific semantic roles. This frame concerns Acts and SVAE.


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 partic-
ipants in certain types of scenario. "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 NOUN identifies the rules or
norms that ought to guide a person's behavior. The FE PROTAGONIST re-
fers 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,
Taher 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 do-
main. Based on data from the British National Corpus and other cor-
pora, FrameNet identifies and describes semantic frames and analyzes
the meanings of words by appealing directly to the frames that under-
lie their meaning. In addition, it studies the syntactic properties of words
by asking how their semantic properties are given syntactic form (Fillmore
7,000 lexical units (LU) (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 follow-
ing steps are taken: (1) characterizing semantically 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 Cor-
pus 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 automat-
ically prepared and stored in the database (for more details, see Fillmore

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.ics.berkeley.edu].
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 con-
sists 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 ac-
cording 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 an-
nnotated example sentence in Figure 7 comply, which is the target (“Tgt”)
evoking the Compliance frame, occurs with the FEs ACT, DIRECTION, 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 sen-
tence occurs in the British National Corpus. FE names are displayed in
terms of subpart 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 Re-
port. 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 re-
spective 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-
| Num | FE|Uset (sort = FE, Compliance, comply, V,) |
|-----|-----------------------------|
| 01  | Act + Degree + comply, V + Norm |
| 02  | Act + comply, V + Norm. |
| 03  | Norm + comply, V + (Protagonist) |
| 04  | Protagonist + comply, V + Degree + Norm |
| 05  | Protagonist + comply, V + Manner + Norm |
| 06  | Protagonist + comply, V + Norm |
| 07  | Protagonist + comply, V + Norm + Taste |
| 08  | State_of_Affairs + comply, V + Norm |
| 09  | State_of_Affairs + comply, V + (Norm) |
| 10  | comply, V + Norm. + (Protagonist) |

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

1. 123614: [Act, Degree]. The last minute addition of the recommendation did not comply with the law and the recommendation would be quashed.

02. : Act + comply, V + Norm

1 123626: The court was told that [Act, Degree] was able to comply with the law.

2. 123755: [Act, Degree]. Spending by public sector organisations has to comply with complex legal regulations.

03. : Norm + comply, V + (Protagonist)

1. 123993: If [Act, Degree], the issue is guilty of an offense, any subsequent contract may be unenforceable and the issue of the advertisement may face criminal charges and/or fines. [Protagonist, CMI]
Comply
Frame: Complier

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

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

<table>
<thead>
<tr>
<th>Frame Element</th>
<th>Number Annotated</th>
<th>Realizations(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Act</td>
<td>(3)</td>
<td>NP:Ext (3)</td>
</tr>
<tr>
<td>Nom</td>
<td>(21)</td>
<td>PP:with(Dep (21)</td>
</tr>
<tr>
<td></td>
<td>DNI:– (1)</td>
<td>NP:Ext (1)</td>
</tr>
<tr>
<td></td>
<td>PP:to(Dep (1)</td>
<td></td>
</tr>
<tr>
<td>Protagonist</td>
<td>(18)</td>
<td>CNE::– (3)</td>
</tr>
<tr>
<td></td>
<td>NP:Ext (15)</td>
<td></td>
</tr>
<tr>
<td>State of Affairs</td>
<td>(2)</td>
<td>NP:Ext (2)</td>
</tr>
</tbody>
</table>

Figure 8. FrameNet entry for comply, Realization table

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 Nom 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: Constructions Null Instantiation (CNI), Definite Null Instantiation (DNI), and Indefinite Null Instantiation (INI). See Fillmore et al. (2003b: 320–321) for more details.
Valence Patterns

These frame elements occur in the following syntactic patterns:

<table>
<thead>
<tr>
<th>Number Annotated</th>
<th>Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 TOTAL</td>
<td>Act</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
<tr>
<td></td>
<td>PP[with]</td>
</tr>
<tr>
<td></td>
<td>Dep</td>
</tr>
<tr>
<td>1 TOTAL</td>
<td>Norm</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
<tr>
<td></td>
<td>PP[with]</td>
</tr>
<tr>
<td></td>
<td>CNI</td>
</tr>
<tr>
<td></td>
<td>_</td>
</tr>
<tr>
<td>16 TOTAL</td>
<td>Norm</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
<tr>
<td></td>
<td>PP[with]</td>
</tr>
<tr>
<td></td>
<td>CNI</td>
</tr>
<tr>
<td></td>
<td>_</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
<tr>
<td>1 TOTAL</td>
<td>Norm</td>
</tr>
<tr>
<td></td>
<td>PP[with]</td>
</tr>
<tr>
<td></td>
<td>Dep</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
<tr>
<td></td>
<td>State_of_Affairs</td>
</tr>
<tr>
<td></td>
<td>DNI</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
<tr>
<td></td>
<td>PP[ho]</td>
</tr>
<tr>
<td></td>
<td>NP</td>
</tr>
<tr>
<td></td>
<td>EX</td>
</tr>
</tbody>
</table>

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, Obara et al. 2003, Boas 2005b, Atkins and Rundell 2000). 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

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 Borchardt 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 Subirá, this volume), or the Romance FrameNet initiative. 14

(2) FrameNets use different types of resources as data pools. That is, besides exploiting a monolingual corpus as is the case with Japanese FrameNet (see Ohara, this volume), projects such as French FrameNet (Pinaï, 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 (Petrick, 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 Krikisaurus (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

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 collocated 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 Medžić's lexical functions (Medžić et al. 1988) in order to identify core frame elements, together with their syntactic role, (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 combinatory 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 multilingual 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 combinational 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.
Schmidt's *The Sictionary – a multilingual lexical resource of football language* directly implements the ideas proposed by Bos in the previous chapter. Schmidt describes the creation of an experimental trilingual 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 is 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 had 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.

Subfields' 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 composition 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 subcorpus creation processes which provide annotators with examples of each possible syntactic configuration in which a lexical item can occur. Part two of Subfields' 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.
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.

Ohira's Frame-based contrastive lexical semantics in Japanese FrameNet: The case of 'risk' and 'kairos' 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.

Petrick'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 (valence) 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 (FEs) 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
Recent trends in multilingual computational lexicography

project is time and labor intensive due to its reliance on the manual crea-
tion of frames as well as the manual annotation of corpus examples.13
The chapter Using FrameNet for the semantic analysis of German: anno-
tation, representation, and automation by Brechtzob 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 (to a limited extent) to describe German predicate-argu-
ment 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 consis-
tency and coverage. This is in contrast to the Berkeley FrameNet, which
focuses on lexicographically relevant examples from the BNC. The chap-
ter details the treatment and annotation of limited compositionality phe-
nomena 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 semi-
tactic parsing. Finally, this chapter addresses some typological differ-
ences (vague items, ambiguity, verb class membership, cross-linguistic para-
phrase modeling, etc.) that arise when applying English-based semantic
frames to the description of German words.

Pilei's chapter on Cross-lingual labeling of semantic predicates and roles:
A low-resource method based on bilingual (t)rait (semantic) analysis
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 les-
tical entries for French can be re-used for other Romance languages such as
Italian, Romanian, Portuguese, and Catalan, which are currently being an-
alyzed by the Romance FrameNet consortium (inspired by MultiSemCor).

The goal of this effort is to (1) create consistent aligned frame-se-
tated multilingual corpus; (2) highlight cross-language regularities, and
structural intra- and extra-typological dissonances; (3) create a semanti-
cally 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 Door 2004, Green et
al. 2004).
languages sub-categorization and collocational mappings; (2) 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 *Intercultural 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 interlingual 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 Vosse and Fillébaum 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 (Vosse 1998) to cover a large number of languages (including lesser-known ones), in the "Global WordNet Grid" (GWG). Vosse and Fillébaum 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-topological differences between multilingual WordNets and multilingual FrameNets, thereby highlighting the different goals of the two approaches.

References

Recent trends in multilingual computational lexicography


Fung, P. and B. Chen
2004

Greber, L. and J. Young
1997

Green, J.
1996

Green, R. and B. Dorr
2004

Greene, R., B. Dorr and P. Resnik
2004

Hamp, B. and H. Feldner
1997

Hasegawa, R.R.K. and G. James
1998

Heid, U.
1996a

Heid, U.
1996b

Heid, U.
1997

Heid, U.
2006

Heid, U., W. Martius and I. Pouch
1991
Possibility and standards for the collocational description of lexi-
cal item, Stuttgart and Amsterdam, EUROTRA-7 Study, Document DOC-9/4.


