The Semantically Based Computer Lexicon HaGenLex

Structure and Technological Environment

Sven Hartrumpf — Hermann Helbig — Rainer Osswald

Applied Computer Science VII, FernUniversität in Hagen, 58084 Hagen, Germany
{Sven.Hartrumpf,Hermann.Helbig,Rainer.Osswald}@fernuni-hagen.de

ABSTRACT. *This article describes the design principles and the technical environment of the computer lexicon HaGenLex, which is a domain independent lexicon for German. Since HaGenLex was developed to support the transformation of natural language expressions into meaning representations, its conception emphasizes semantic aspects. The representation of semantic information in HaGenLex is based on the semantic network formalism MultiNet. The formal description of HaGenLex employs a fairly standard typed feature architecture, which is extended by a class-based inheritance formalism that allows default specifications. The development and maintenance of HaGenLex is supported by several software tools including a lexicon browser and editor as part of a user-friendly and powerful workbench for the lexicographer. Moreover, HaGenLex provides various interfaces to corpora as well as to other lexical databases, which are also used for lexical validation and cross-checking.*

RÉSUMÉ. *Cet article décrit les principes de conception et l'environnement technique de HaGenLex, un dictionnaire électronique général de l'allemand. Comme HaGenLex a été développé dans le but d'assister la conversion de formes linguistiques en représentations sémantiques, sa conception met l'accent sur les aspects sémantiques. Leur représentation se fonde sur le formalisme de réseau sémantique MultiNet. HaGenLex utilise une architecture de traits typés classique. Celle-ci est complétée par un formalisme d'héritage de classes permettant la spécification de valeurs par défaut. Le développement et la maintenance d'HaGenLex sont supportés par une série d'outils (dont un fouineur et un éditeur) intégrés dans un atelier convivial et puissant destiné aux lexicographes. Par ailleurs, HaGenLex fournit plusieurs interfaces permettant l'accès aux informations contenues dans des corpus ou d'autres bases de données lexicales, qui sont aussi utilisés pour la validation et la vérification des données.*

KEYWORDS: *inheritance-based lexicon, lexical semantics, empirical lexicon validation, lexicon tools, corpus tools.*

MOTS-CLÉS : *dictionnaire fondé sur la notion d'héritage, sémantique lexicale, validation empirique des données lexicales, outils lexicaux, outils de corpus.*

1. Introduction

HaGenLex (**Ha**gen **Ge**rma**n Lex**icon) is a domain independent computer lexicon for German. It has been developed to support the automatic transformation of natural language expressions into semantic representations that are based on the so-called *MultiNet paradigm* ([HEL 01]). Providing detailed semantic information in terms of MultiNet specifications is thus an essential requirement for HaGenLex entries.

The MultiNet (*Multilayered Extended Semantic Networks*) paradigm serves as a framework for natural language semantics and knowledge representation in general. "Meaning" or "knowledge" is represented in form of *semantic networks*, which are labeled directed (hyper)graphs whose nodes represent concepts and whose edges represent semantic relations. Intuitively, a concept is the mental picture of an entity in a real or cognitive world. Within the HaGenLex-MultiNet approach, it is assumed that each reading of a (content) word uniquely corresponds to a concept. Such concepts are called *lexicalized*. Consequently, there is a one-to-one correspondence between lexemes and lexicalized concepts of the language.[1]

The lexical information of HaGenLex is intended to be neutral with respect to any specific theory of grammar. It should hence be possible without too much effort to adapt HaGenLex, viewed as a lexical resource, to specific grammatical frameworks like LFG ([BRE 01]) or HPSG ([POL 94]) – in particular, if the MultiNet paradigm is adopted as the underlying semantic framework.[2] In current natural language processing (NLP) applications of HaGenLex, the syntactic analysis is based on so-called *Word Class Functions* ([HEL 97]). This approach has been implemented in the syntactico-semantic parser WOCADI (WOrd ClAss based DIsambiguating parser, [HAR 03]), which is employed, for instance, in natural language interfaces to databases ([HEL 00]) and for information retrieval on the Internet ([LEV 02]).

The core of HaGenLex consists of lexical material that has been manually compiled by means of a powerful workbench for the lexicographer. In addition, HaGenLex is being continually extended on the basis of both, corpus data ([HAR 03]) and publicly available electronic dictionaries like GermaNet ([KUN 01]) and CELEX ([BAA 95]).[3] (The latter two linguistic databases have been used to supplement HaGenLex with lexical semantic relations and morphological information, respectively, as well as to automatically build underspecified lexica for fallback strategies.) With its current stock of about 20,000 fully syntactically and semantically described lexemes, plus 50,000 lexemes bearing only morpho-syntactic characterizations, and more than

1. Though HaGenLex is a monolingual lexicon for German, we claim that its semantic basis can be employed for multilingual lexica as well.
2. The usual commitment of HPSG to Situation Semantics seems to be independent of HPSG as a grammatical theory.
3. Unfortunately, no machine readable dictionaries for German are available hitherto that would be suitable for the semi-automatic acquisition of syntactic and semantic information in style of the ACQUILEX project ([COP 93, SAN 93]).

200,000 proper nouns, HaGenLex can compete with many other computer lexica for German used in NLP tasks.[4]

A central feature of the HaGenLex-MultiNet approach is its technological embedding into a system of NLP modules and software tools which support the acquisition of all aspects of knowledge connected with NLP systems. This includes a workbench for the lexicographer (Section 5.1), a workbench for the knowledge engineer ([GNÖ 02]), and several tools for acquiring linguistic knowledge from corpora ([STA 02]).

As for other conceptions of semantically based computer lexica, we only mention Pustejovsky's Generative Lexicon ([PUS 95]). (A detailed comparison with other approaches is beyond the scope of this article.) Due to its focus on lexical semantics, it seems fair to say that the representational means for semantic information used in the Generative Lexicon are not intended to serve as a general knowledge representation formalism. In the case of HaGenLex, in contrast, there is a transparent integration of lexical semantic information in all modules based on the MultiNet paradigm. For example, the process of inferential answer finding over MultiNet knowledge bases has direct access to lexical semantic information. This transparency is one of the design criteria of MultiNet, called the *interoperability criterion* ([HEL 01, HEL 02]).

The rest of the article is organized as follows: Section 2 outlines the main ideas underlying the conception of HaGenLex and gives an overview of its content. In Section 3, a brief exposition of the MultiNet paradigm is given. Section 4 describes the internal representation of HaGenLex entries by typed feature structures and the design of HaGenLex as an inheritance-based lexicon. Section 5 gives an overview of several software tools that support the development, maintenance, and application of HaGenLex. In Section 6, the central issue of validating the quality of HaGenLex entries is addressed from various perspectives. The concluding Section 7 points out some possible extensions and improvements of the HaGenLex project.

2. Lexical Structure and Content

2.1. *Semantic Categorization*

The semantic information in HaGenLex is based on the representational means provided by the MultiNet paradigm. The "upper ontology" of MultiNet consists of a tree-shaped hierarchy of 45 *ontological sorts*; see Appendix, Table 3. Since HaGenLex assumes a one-to-one correspondence between word meanings and lexicalized concepts, each lexical entry constrains the ontological sort of the respective concept.

The classification by ontological sorts, however, is not fine-grained enough to adequately describe *selectional restrictions* and hence to support *disambiguation* during

4. The VERBMOBIL project ([WAH 00]), for example, uses a much smaller lexicon, although one must concede that the VERBMOBIL task of speech-to-speech translation involves many aspects not addressed by the HaGenLex-MultiNet system.

syntactico-semantic analysis. For example, there is no ontological sort to distinguish between the agents admissible to the German verbs "*essen*"[5] and "*fressen*", which mean "*eat*" with a human and an animal agent, respectively. It is possible to express such restrictions within MultiNet by specifying that the entity realized by the first argument of the verb is subordinate to the concept *Mensch* (Eng. *human*) or *Tier* (Eng. *animal*), respectively. The lexical entry of "*essen*" would then contain the MultiNet specification (SUB *x1 Mensch*) as part of the value of the feature NET; cf. Section 2.3.

On the other hand, using the full descriptive repertory of MultiNet to express the selectional restrictions of lexemes is disadvantageous from a practical point of view. Firstly, to make use of this kind of information the syntactico-semantic analyzer would need to invoke the full inference mechanism of MultiNet, which is too time-consuming. Secondly, the lexicographer's decisions in specifying such restrictions are rather inefficient and unconstrained. For these two reasons, HaGenLex employs a fixed set of 16 binary *semantic features* (see Appendix, Table 1) for the semantic classification of concepts and the specification of selectional restrictions. Checking selectional restrictions then reduces to a simple unification of feature vectors.

Notice that the values of the semantic features listed in Table 1 are not independent of each other – for instance, [HUMAN +] implies [ANIMATE +]. HaGenLex takes this into account by a predefined set of semantically coherent (sort and) feature combinations. Such a combination of an ontological sort with semantic features will be referred to as a *semantic sort*. To give an example, one of the semantic sorts of HaGenLex is *con-info* (concrete information object), which is defined by the ontological sort *d* (discrete object) and the semantic features [ANIMATE −], [ARTIF +], [INFO +], and [MOVABLE +], besides others. The semantic sort *con-info* is used for lexemes like "*picture*" and "*newspaper*".

As for *lexical ambiguity*, HaGenLex allows to distinguish between *homographs*, *polysemes*, and so-called *meaning molecules*. A meaning molecule is a regular polyseme whose different *meaning facets* can be simultaneously referred to in a single context.[6] The lexeme "*school*", which can refer to a building and an institution, serves as a standard example – witness the following sentence: "*The school across the street has a good reputation.*" Technically, the meaning facets of a molecule are represented in a single HaGenLex entry by a disjunction of semantic sorts (as value of the lexical feature ENTITY) and an indication of molecularity (by the value + of the lexical feature MOLEC); cf. Section 4.1. Another type of regular polysemy worth mentioning is *verbal alternation* ([LEV 93]). An adequate treatment of alternations for German verbs by lexical rules on the basis of MultiNet semantics is a current project to improve lexical representation in HaGenLex.

---

5. We use the typographic convention that e.g. *"eat"* is the word reading whereas *eat* is the respective concept. Here (and at some other places in this article) we do not distinguish between different readings. But notice that in HaGenLex, different readings of the same word are systematically distinguished by numerical indices.

6. A similar subclassifi cation of regular polysemy is proposed in [BUI 98, Sect. 3.2].

AGT	OBJ	MCONT
np / nom	*np / acc*	*"über"-pp / acc*
	optional	optional

Figure 1. *Case frame sketch for the verb "informieren" (Eng. "to inform")*

2.2. *Valency and Case Frames*

Conceptual entities denoted by lexemes often co-occur with other entities in a situational description. A situation denoted by a verb, for instance, typically bears certain relations to other entities that participate in the situation in question. In HaGenLex, this information is represented by the MultiNet relations, or *cognitive roles*, the denoted situation bears to its participants.[7] (The set of cognitive roles provided by MultiNet is presented in Table 2 of the Appendix.) As mentioned in Section 2.1, a HaGenLex entry may also specify *selectional restrictions* regarding the ontological sort and the semantic features of these participants.

The list of cognitive roles, i.e. the *semantic case frame* of the lexeme, is supplemented by the possible syntactic realizations of the corresponding arguments, i.e. by a *syntactic case frame*. The architecture of HaGenLex allows to capture regularities between the syntactic realization of an argument and its cognitive role by means of so-called *selectional classes*; see Section 4.2. Whereas participants in the valency list of a lexeme are regarded as *semantically obligatory*, their syntactic realization may be optional. Figure 1 shows a sketch of the information to be stored in the case frame of the verb "*informieren*". (The complete entry is presented in Figures 6 and 8 below.)

Although *free adjuncts* are not determined by a lexeme, the latter can impose certain restrictions on the former. For example, a modification by duration adverbials is not always possible but depends on the aspectual class of the verb. Within a HaGenLex entry, adjuncts can be restricted by explicitly listing (in the lexical feature COMPAT-R) all MultiNet relations that are compatible with the corresponding concept.

2.3. *Lexical Semantic Relations*

The MultiNet formalism provides a set of standard *lexical semantic relations* including SYNO (*synonymy*), SUB (*subordination*, *hyponymy*), and ANTO (*antonymy*) (which splits further into COMPL (*complementarity*), CONTR (*contrariness*), and CNVS (*conversion*)). These relations are employed in HaGenLex to represent semantic relations between lexicalized concepts and hence between lexemes.

The word meanings of HaGenLex are systematically linked (by the lexical feature G-ID) to those of the GermaNet database ([KUN 01]). It is thus possible to automat-

7. A cognitive role closely resembles what is often called a *thematic role*.

ically project the semantic relations of GermaNet onto HaGenLex. Besides being of use for lexical inference tasks, the linking to GermaNet also allows to validate the consistency of HaGenLex relative to GermaNet; for example, synonymous entries should be of identical or closely related semantic sorts; see Section 6.

In MultiNet, the concepts corresponding to verbs and deverbal nouns are subsumed under different ontological sorts, *situation* [*si*] in the first case and *situational object* [*abs*] in the second (cf. Appendix, Table 3). Likewise, concepts expressed by adjectives and deadjectival nouns differ with respect to their ontological sort, which is *quality* [*ql*] and *attribute* [*at*], respectively. MultiNet allows to capture the systematic relationship between concepts that parallels the *derivational* interdependence between lexemes by several *change-of-sort relations*. The relation CHPA, for instance, indicates the change from a property to an attribute, e.g., from *friendly* to *friendliness*.[8]

Whereas the categorization of lexemes by semantic sorts and case frames is fully integrated into the typed feature structure representation of HaGenLex entries (cf. Section 4.1), additional semantic specifications can be encoded as value of the feature NET, which allows to include arbitrary MultiNet expressions into lexical entries.[9] For example, the expression (SYNO c *produzieren*) as part of the NET value of the lexeme "*herstellen*" (Eng. "*to produce*") means that *herstellen* is synonymous to *produzieren*, whereas (CHEA c *Herstellung*) says that *Herstellung* is the abstract object corresponding to the situation *herstellen*. (The symbol c by convention stands for the concept of the given entry.) Notice that this method of specifying MultiNet expressions also allows to include more general *meaning postulates* into HaGenLex.

2.4. *Further Aspects*

The *morphological* specification of a HaGenLex lexeme is essentially determined by its *inflectional paradigm*. HaGenLex employs the DUDEN classification of inflectional types ([DRO 95]). A considerable part of the inflectional information of HaGenLex entries has been semi-automatically extracted from the CELEX database ([BAA 95]).

Some *function words* like prepositions and conjunctions must be described with semantic information if one wants to follow the semantic orientation of HaGenLex. A given preposition may give rise to many different interpretations depending on the semantics of the sister NP and the semantics of candidate mother constituents. HaGenLex currently contains 285 such preposition readings for 122 prepositions; [HAR 99] describes how these rules can be used in a hybrid PP disambiguation method.

Although *proper names* are typically not considered part of the main lexicon, they are a vital element if one wants to achieve wide coverage for semantically oriented

8. The letters P and A in CHPA signify property and abstract object, respectively. Similar for CHPE, CHEA, CHSA, and CHPS, with S and E signifying state and event, respectively.
9. See [HEL 01] and [GNÖ 02] for more about formal specifications of MultiNet networks.

NLP systems. We currently maintain around 40 different name lexica. Some of them are (semi-)automatically derived from extant ones, some (smaller ones) are manually created. Each name lexicon defines to which semantic class an object named by an entry of the name lexicon belongs.

The coverage of the HaGenLex system is further extended by a powerful *compound analysis* module for analyzing compound nouns, adjectives, and (prefix) verbs, which makes use of HaGenLex and additional name lexica. If possible, the compound receives semantic information from its components. The compound analysis module of HaGenLex is just one of several fallback strategies in case a lexeme is missing in the HaGenLex core lexicon. Another one is to fall back on the supplementary morpho-syntactic lexicon mentioned in the introduction, which is based on other lexical databases like CELEX.

3. The MultiNet Paradigm

3.1. *Synopsis*

The MultiNet paradigm is a semantic interlingua fulfilling the *universality requirement* in three aspects: it is language independent, has a broad coverage of the semantic phenomena of natural language, and is independent of any application domain. MultiNet has been developed along the line of semantic networks starting with the work of Quillian ([QUI 68]). One of the design principles of MultiNet is the *interoperability criterion* ([HEL 01, p. 7]) requiring that a knowledge representation system to be used for natural language processing must be appropriate to describe lexical knowledge as well as world knowledge, linguistic knowledge as well as inferential knowledge. With regard to HaGenLex, the expressional means of MultiNet provide the semantic backbone because they are used to specify the meaning structure of lexemes. Since there is a one-to-one correspondence between lexicalized concepts and lexemes in our approach, all statements about the semantic representation of concepts in MultiNet have an immediate effect on the characterization of lexemes of HaGenLex. It must be emphasized that this article deals only with those parts of MultiNet that are relevant to the lexicon; for further details the reader is invited to consult [HEL 01].

MultiNet networks are hypergraphs whose nodes represent concepts, and the arcs between the nodes represent relations and functions establishing a semantic connection between these nodes. In contrast to other network formalisms (e.g. KL-ONE, [BRA 85]), the arcs are labeled by elements of a predefined set of relations and functions; cf. Section 3.2 below. These relations and functions can be represented as nodes of a second semantic network at a meta level which are connected with axioms and inference rules describing the logical properties of these meta level constructs. Another aspect of MultiNet distinguishing it from more simple semantic networks is the inner structure of nodes: each node is embedded in a multidimensional space of so-called layer attributes that encode e.g. genericity and definiteness information.

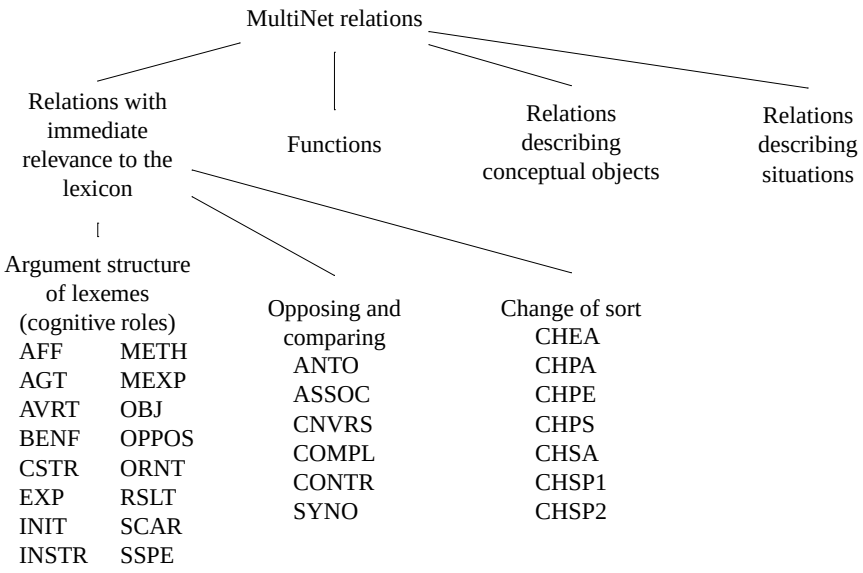

Figure 2. *The lexically relevant relations of MultiNet*

3.2. *Ontological Sorts, Layer Attributes, and Semantic Relations*

MultiNet distinguishes 45 ontological sorts, which are arranged in a tree-shaped hierarchy (see Appendix, Table 3). Among other things, the sorts are important for the definition of the *signatures* of the semantic relations and functions (see below). Apart from ontological sorts, MultiNet concepts (and therefore also HaGenLex lexemes) are classified with respect to the seven *layer attributes* FACT, GENER, QUANT, REFER, CARD, ETYPE, and VARIA. The values of these attributes indicate, e.g., the facticity or existence of an entity, its determination of reference, and its type of extensionality; see Appendix, Table 4, for a more detailed description. In the lexicon, for instance, the type of extensionality helps to characterize collective nouns. (Layer specifications are especially useful for imposing semantic constraints on the combination of determiners, quantificators, and content words; see [HAR 02].)

MultiNet provides a repertory of about 140 semantic relations and functions to describe the connections between concepts (i.e. nodes of the semantic network). In principle, all of these formal constructs can also be used to describe lexical entries (lexemes), because through the lexical feature NET mentioned in Section 2.3, all lexemes can be embedded in a larger semantic network to describe their meaning. Figure 2 highlights those MultiNet relations that are particularly important for the semantic characterization of lexemes, namely cognitive roles to characterize the semantic

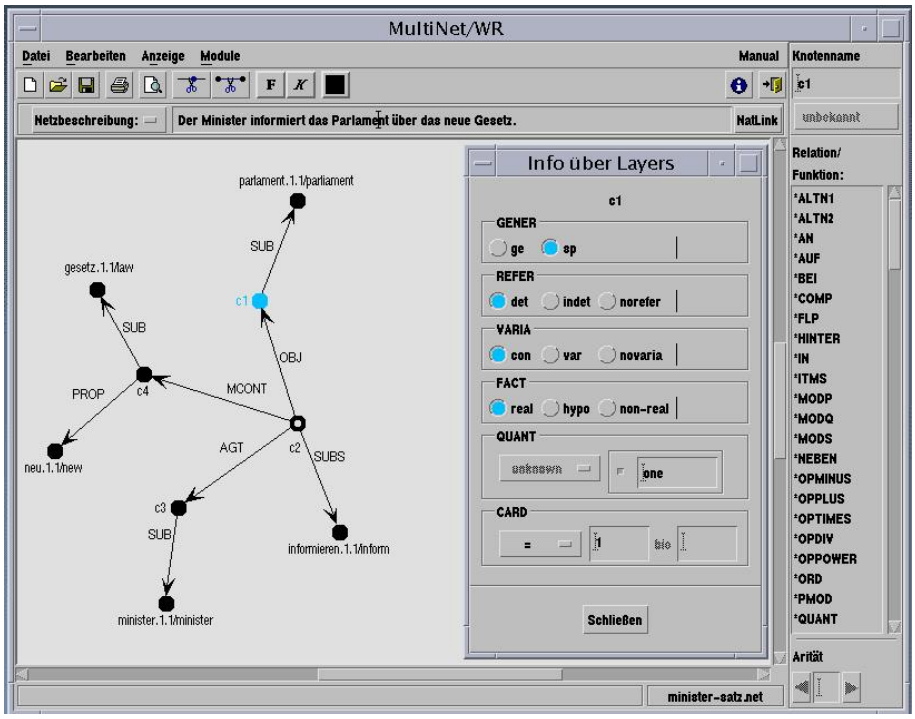

Figure 3. *Semantic representation of the German sentence "Der Minister informiert das Parlament über das neue Gesetz." (Eng. "The minister informs the parliament about the new law.")*

case frame of content words (see also Appendix, Table 2), lexical semantic relations for comparing and opposing lexicalized concepts, and change-of-sort relations.

A simple illustration is given in Figure 3, which shows the semantic representation of the sentence "*Der Minister informiert das Parlament über das neue Gesetz.*" (Eng. "*The minister informs the parliament about the new law.*"). The presentation in the Figure makes use of a graphical MultiNet editor ([GNÖ 02]). It displays the layer information of the concept node *c1* that corresponds to the expression "*das Parlament*"; this concept is subsumed by the generic concept *Parlament.1.1.*

It is essential for both the lexicographer and the knowledge engineer working within the MultiNet paradigm to have a detailed explication of all semantic primitives at his disposal. To this end, each MultiNet relation is described by its signature, a verbal definition, a mnemonic hint, one or more question patterns, and various comments and examples that explain the proper application of the relation in question. Figure 4 shows the (abbreviated) description of the relation MCONT that connects a mental or informational process and the content of that process.

MCONT : $[si \cup o] \times [si \cup o]$

Definition: The relation MCONT allows to specify the content of a mental or informational process. MCONT is also used as shorthand for the content of the *result* of such a process.

Mnemonics: mental content
(MCONT x y) – x is characterized by the informational or mental content y

Question patterns: What [do] ... {think | reason | dream | assume | ... }?
What [do] ... {say | convey | contain | hear | tell | ... }?
What [do] ... {learn | come to know | experience | ... }?
What [do] ... {believe | remember | ... }?
{About what | Of what} [do] ... {speak | write | think | ... }?

Comments: By default, the second argument of MCONT is unspecified with respect to validity, if it is a situation, or existence, if it is an object. Without additional information available, this argument is thus marked as hypothetical by means of the layer specification [FACT *hypo*]. Syntactically, the second argument of MCONT is often realized as a subordinate clause or an infinitive construction. In general, the second argument of MCONT is not independent of the validity or existence of the first argument. In order to explicitly express this independence, the relation OBJ can be used in addition to MCONT. Examples:

"The mathematician believed that [*he had found a proof*] $^{\text{MCONT}_{arg2}}$."
"The girl dreams of [*her boyfriend*] $^{\text{MCONT+OBJ}_{arg2}}$."
"Peter is engrossed with the idea of [*going on a holiday trip*] $^{\text{MCONT}_{arg2}}$."
"A book about [*the mammoth cave*] $^{\text{MCONT}_{arg2}}$ ..."
"The notification of [*his friend's death*] $^{\text{MCONT+OBJ}_{arg2}}$ ..."

Figure 4. *Definition of the semantic relation MCONT (abbreviated)*

4. Representation and Implementation

4.1. *Feature Architecture*

The internal representation of HaGenLex entries is based on a *typed feature structure* formalism along the lines of [CAR 92]. In addition to lists and disjunctions, we also allow sets of atomic types as feature values. Currently, our implementation does not support path identities, which is no severe restriction because the HaGenLex feature architecture is designed for lexical information only and not for phrasal constraints (as e.g. in HPSG, [POL 94]).[10]

The *type hierarchy* of HaGenLex has the form of a taxonomic tree. In particular, the immediate subtypes of any type are pairwise incompatible. Besides fairly standard types like *case*, with subtypes *nom*, *gen*, *dat*, and *acc*, the HaGenLex type hierarchy also includes the ontological sorts and the lexically relevant semantic relations of MultiNet. As usual in typed feature approaches, a feature of the HaGenLex feature

10. See [HAR 03, HEL 97] for the grammar model used in the HaGenLex-MultiNet system.

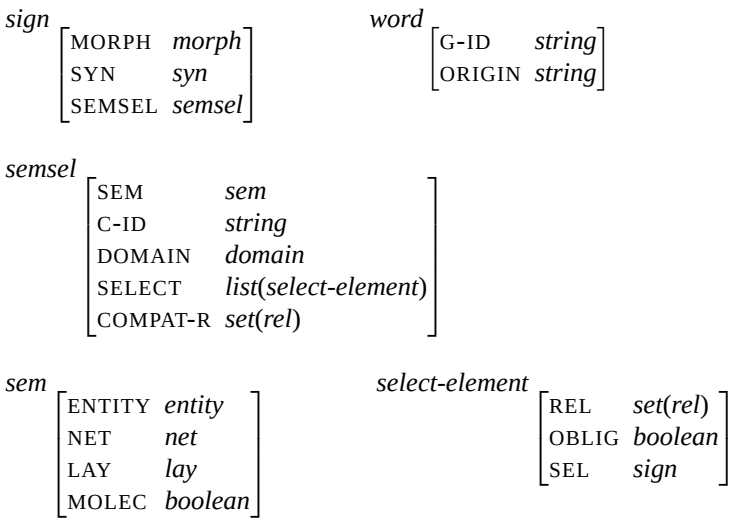

Figure 5. *Examples of feature declarations used in HaGenLex*

architecture is *appropriate* only in structures of a certain type. The feature MORPH, for instance, is only appropriate in structures of type *sign*. Moreover, the value of a feature is restricted by the type the feature is appropriate to. It is furthermore presumed that a type inherits all features appropriate to its supertype. So the feature MORPH is appropriate to the type *word* because *word* is a subtype of *sign*. A (non-redundant) list of the appropriate features and their respective values for a given type is known as the *feature declaration* of that type.

Figure 5 displays three instructive examples of HaGenLex feature declarations (with minor simplifications). Since the feature structure representation of a lexical entry is of type *word* and *word* is a subtype of *sign*, the topmost feature level of such a structure is determined by the feature declarations of *word* and *sign*. The value of the feature SEMSEL is a structure of type *semsel*, whose topmost feature level is given by the declaration of *semsel*; feature structures of this type represent the semantics and the valency of a lexeme. Valency in turn is encoded by a list of structures of type *select-element*, each of which characterizes a complement by a set of semantic relations (REL), its syntactic obligatoriness (OBLIG), and its description by a structure of type *sign* (SEL);[11] cf. the feature declaration on the lower right of Figure 5. Structures of type *sem*, finally, represent the semantics of a lexical sign by its semantic sort (ENTITY), additional MultiNet expressions (NET), layer information (LAY), and molecularity type (MOLEC). For an actual example from HaGenLex, see the feature structure representation of the verb "*informieren*" shown in Figure 6. It is worth

11. Morphological information about complements is needed e.g. in the case of support verb constructions.

mentioning that Figure 6 has been automatically generated by LIA (Section 5.1) and IBL inference (Section 4.2) from the compact IBL representation of "*informieren*" as presented in Figure 8.[12]

Notice that a semantic sort is represented by a feature structure of type *entity*. For structures of this type, the semantic features described in Section 2.1 are appropriate, with possible values *boolean*, $+$, and $-$; in addition, the feature SORT is appropriate, whose possible value are the ontological sorts of MultiNet. In other words, the definition of semantic sorts as combinations of ontological sorts and semantic features is modeled by feature declarations.

4.2. *The IBL Formalism*

The lexicon representation of a lexeme as a typed feature structure, where most feature paths bear an explicit value (see preceding section), is well suited for using the information contained in the lexicon, for example in natural language parsers. For creating and maintaining the information, matters differ. The expanded lexicon representation contains many redundancies, and regularities are not explicit. To improve this situation, the concept of *classes* is introduced from the IBL (Inheritance-Based Lexicon) formalism so that the lexicon becomes an inheritance-based one.[13] Please note that although IBL entries are an important step for easing the creation and maintenance of lexical entries, we still consider a graphical and intelligent user interface as a second vital step in this direction. Such a tool, LIA, is described in Section 5.1.

A class is a named collection of attribute-value constraints that describes a typically underspecified typed feature structure. For example, the class *agt-select* shown in Figure 7 captures the regularity between the semantic complement relation AGT from MultiNet and its syntactic realizations: the role AGT can be realized (in the context of verb complements, inherited from the nonlexical class *vselect*) as a nominative NP (*np-nom-syn*) or a "*mit*"-PP (*mit-dat-pp-syn*); a question mark precedes the default value (?*np-nom-syn*); in addition, such a complement is always syntactically obligatory (OBLIG $+$). A class can inherit from other classes at top level or at specific path locations. The latter case may be called *locating inheritance*.[14]

A *lexical class* is the IBL view of a lexical entry (or lexeme), that is, a lexeme is represented – with no or low redundancy – as a lexical class that inherits from one or more superclasses. An example entry is shown in Figure 8. *Nonlexical classes* do not describe single lexemes but are used in several lexical classes or other nonlexical classes. In HaGenLex, nonlexical classes also serve to avoid errors when one writes lexical classes (manually or semi-automatically by the use of LIA): many nonlexical

12. The presented verb reading is assumed to semantically restrict only the first two arguments.
13. A Prolog-based predecessor implementation is described by [HAR 96, HAR 94]; the ELU formalism ([RUS 92]) bears some similarities.
14. In order to force the encapsulation and explicit naming of all important information bundles, IBL does not allow the inheritance of parts of a class, in contrast e.g. to DATR.

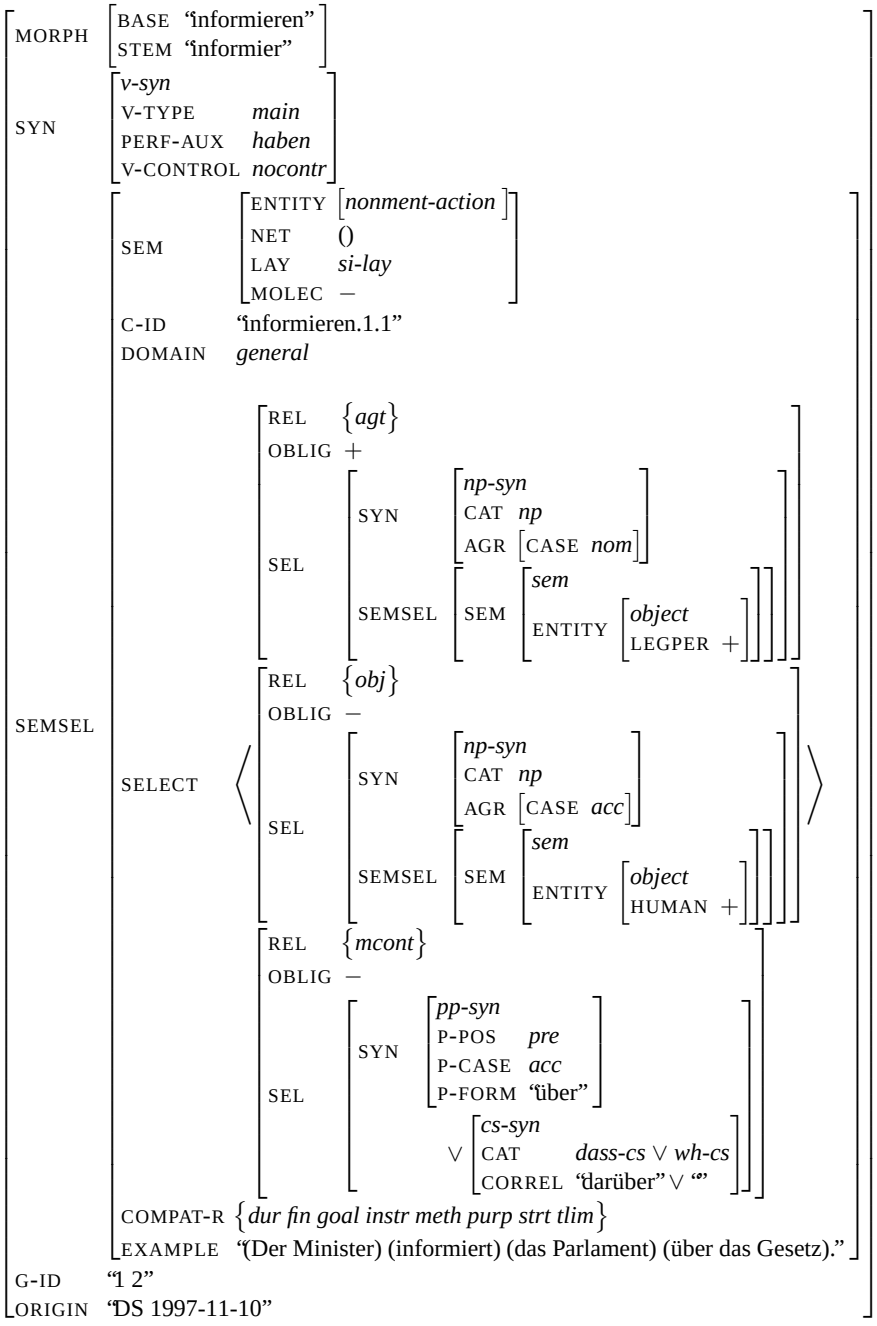

Figure 6. *Feature structure representation of the HaGenLex entry for "informieren" (expansion of the lexical class in Figure 8)*

```
agt-select [
   vselect
   rel {agt}
   oblig +
   sel syn (np-nom-syn mit-dat-pp-syn) ?np-nom-syn]
```

Figure 7. *The nonlexical class agt-select*

```
"informieren.1.1" [
   verb
   semsel [
      v-nonment-action
      select <
         [agt-select
            sel semsel sem entity legper +]
         [obj-action-select
            oblig -
            sel semsel sem entity human +]
         [mcont-select
            oblig -
            sel syn (ueber-acc-pp-syn darueber-dass-syn wh-syn)] >
      compat-r {dur tlim}
      example "(Der Minister) (informiert) (das Parlament) (über das Gesetz)."]
   g-id "1 2"
   origin "DS 1997-11-10"]
```

Figure 8. *A lexical class for one of the readings of the verb "informieren"*

classes like the ones for complements specify the range of possible values (e.g. syntactic category); a violation of these range restrictions within a lexical class can then be detected by expanding this class (i.e. by making explicit all inherited information).

5. Technological Environment and Applications

5.1. *The Lexicographer's Workbench*

In order to build a reasonable large lexical database, it is indispensable to have an effective and easy-to-handle user interface. For, firstly, it would be a tedious task to specify lexical entries by explicitly listing the appropriate feature, type, and class specifications, and, secondly, only experts familiar with the internal feature architecture and class hierarchy of HaGenLex would be able to perform this task.

The lexicographer's workbench LIA is a software tool that allows to write and edit HaGenLex entries without any knowledge about the internal representation of lexemes

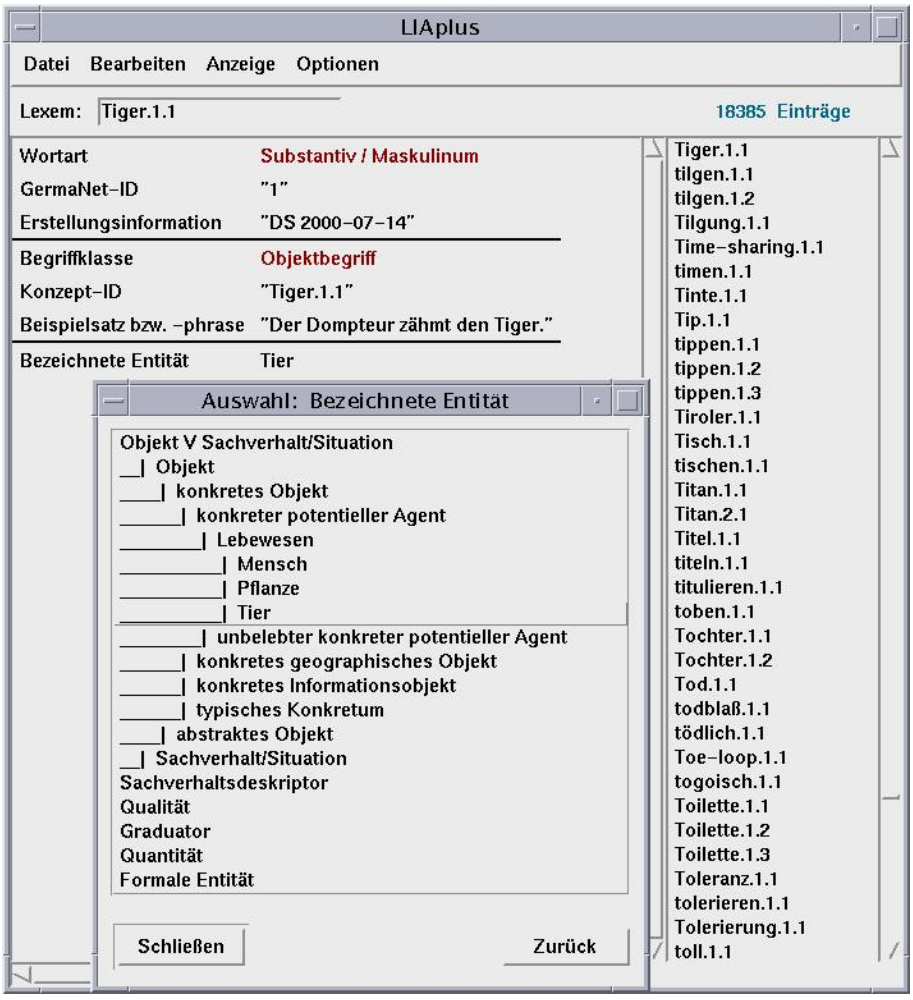

Figure 9. *Selecting the semantic type of the noun "Tiger" by means of the lexicographer's workbench LIA*

by features, types, and classes.[15] To this end, LIA explicitly lists possible choices, say, of the semantic type of a noun or the class of a verbal complement, in terms of natural language paraphrases. The screenshot displayed in Figure 9, e.g., shows the scenario of selecting the semantic type for the main reading of the German noun "*Tiger*" (Eng. "*tiger*"), where the internal representation of the type, which is hidden from the user, is presented by LIA as "*Tier*" (Eng. "*animal*").

15. See [SCH 98] for a detailed description of an earlier version of LIA.

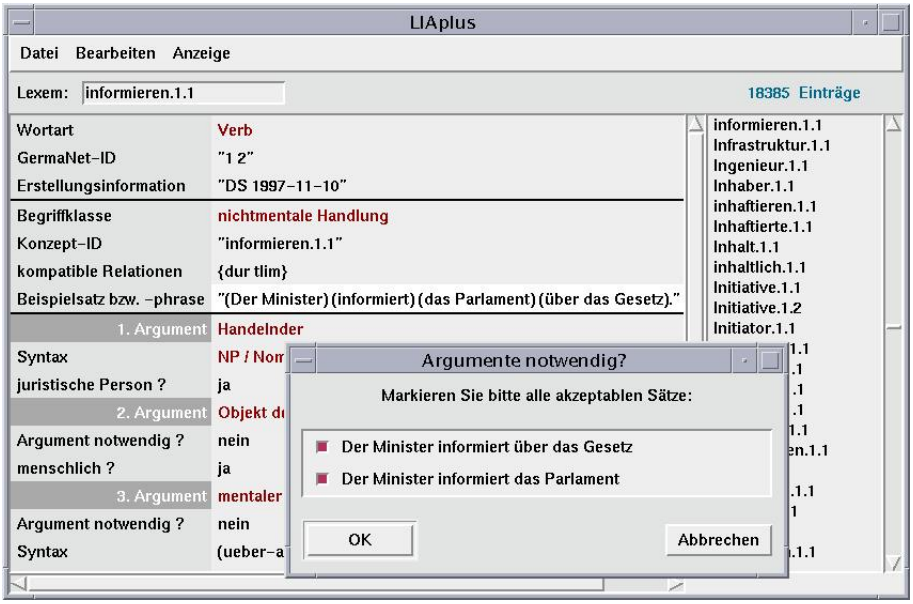

Figure 10. *LIA queries the complement status by means of acceptability decisions*

The philosophy of LIA is to query information in a way that takes advantage of the linguistic intuitions of the native speaker. So the typical user of LIA needs only a moderate linguistic background. For instance, LIA supports the decision whether the complements of a verb are obligatory or optional by presenting example sentences with one complement omitted. The user is then asked to give acceptability judgments. The screenshot of Figure 10 illustrates this procedure for the verb "*informieren*" (Eng. "*to inform*"), where both complements are classified as optional.

LIA provides several interfaces to other software components of the HaGenLex-MultiNet system. For example, recall from Section 2 that HaGenLex lexemes may contain additional semantic specifications in terms of arbitrary MultiNet expressions as values of the feature NET. To facilitate editing of these expressions, LIA allows to automatically load them into the graphical MultiNet editor and to save the modified result as value of NET. Another application of this interface, now in combination with the syntactico-semantic analyzer WOCADI (see Section 5.2), is illustrated by the screenshot of Figure 11. It shows how LIA allows to analyze the example phrase(s) of an entry by invoking WOCADI. The lexicographer can thus immediately check whether an entry is correctly specified with respect to the example context. To supply this test procedure with a sufficiently representative set of phrases, we are currently developing an interface to the corpus tools described in Section 5.2 below. In particular, the corpus interface will support the lexicographer's decisions, say, concerning the number of obligatory complements, by actual occurrences in corpora.

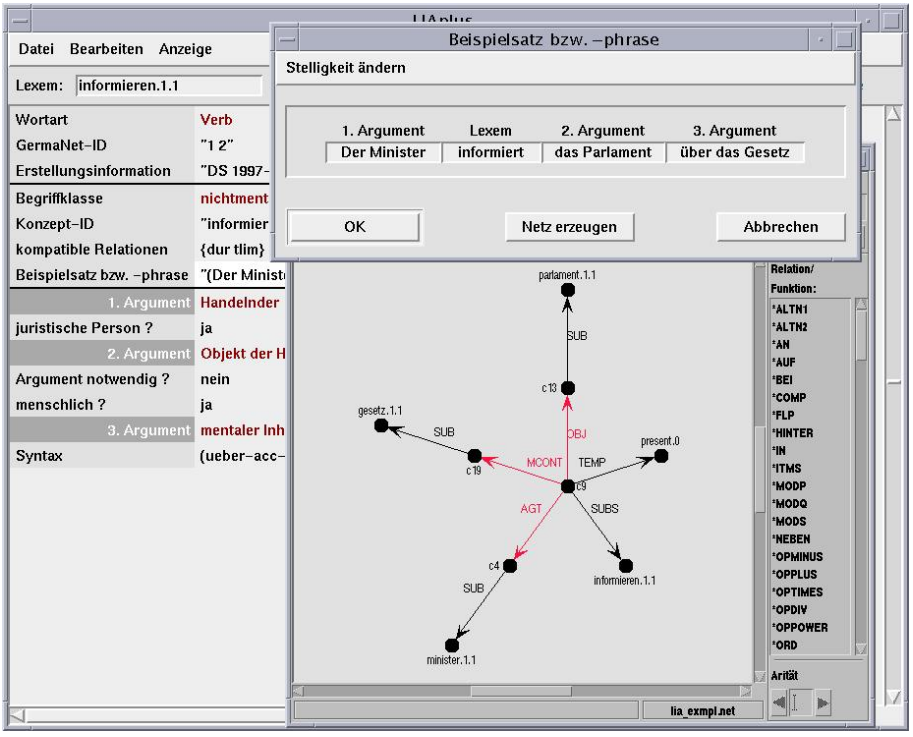

Figure 11. *Integrated semantic analysis of an example sentence in the lexicon*

The current implementation of LIA uses a two-level architecture (in contrast to an earlier implementation described in [SCH 98]). The front-end of LIA is implemented in Tcl/Tk; it controls the graphical user interface, hides the internal representation, and sets up the communication with a back-end Scheme application that realizes the actual inferences triggered by user actions. The inference machine of LIA is based on the feature declarations and class definitions described in Section 4. In addition, there are LIA specific lexical rules to speed up the editing process by default inferences.

5.2. *Parser and Corpus Tools*

The WOCADI (WOrd ClAss based DIsambiguating) parser is one central NLP program that relies on HaGenLex and produces semantic networks of the MultiNet formalism for German phrases, clauses, sentences, and texts. As such, it is an ideal source of feedback for the lexicon. In Section 6, some specialized and integrated uses of the parser for the creation and maintenance of HaGenLex are described.

The lexicographer working with HaGenLex can use corpora in several ways to investigate readings and their linguistic environments. First, there is a word form oriented search mainly intended for new entries; second, there is concept-oriented search mainly intended for refining existing entries or adding related readings. The word form oriented method is realized by a client-server architecture. It is basically a KWIC (Key Word In Context) approach for locally available corpora (see Section 6). The server provides efficient access to word forms and their linguistic contexts by way of large indices. The client, which can work via HTML forms or a command line interface, passes user queries to the server and presents the returned results.

The second search method, which is realized by a similar architecture, allows to search for concept IDs in the parsed corpus instead of word forms in the raw corpus. Thus, recall is increased (compared to the word form oriented search) by way of lemmatization, and precision is increased by way of the (disambiguating) parser WOCADI. A result can be returned in several formats, including the MultiNet representation of the enclosing sentence. So, one could call this a KWIN (Key Word In Network) approach for parsed corpora.

The third and most powerful corpus tool is called NetArt. It provides a query language for semantic and syntactic information contained in parsed corpora. For example, one can search for semantic subnetwork patterns containing variables and/or for syntactic dependency configurations. The output format can easily be tailored to suit other tools, e.g. machine learning tools.

6. Validation and Quality Assurance

Cross-checking by generalized indices. Cross-checking, i.e. the improvement of correctness and consistency inside the lexicon, is supported by two indexing tools.[16] First, LIA provides powerful, fully interactive, and efficient indexing: By clicking on feature values one can incrementally extract the set of lexemes carrying the chosen feature values. For example, mouse clicks on the three complement relations of a ternary verb instantly list all verbs in the lexicon that share the same semantic case frame (they appear in the right frame of LIA, see Figure 10). In this way, even the first step of lexical work, the creation of new entries, can be supported by cross-checking and cross-comparison. Second, for complex indices that group lexemes together depending on Boolean combinations of attribute-value pairs, there exists an index program that generates indices using hypertext capabilities and PDF or HTML as output format. Both index approaches have their own justification and merits. While the first is ideal for day to day (i.e. lexeme to lexeme) practice, the latter is more adequate for rarer investigations hunting down specific representation problems.

Parsing example sentences. Each entry in HaGenLex contains one or more example sentences (or phrases) that illustrate the represented reading (feature EXAMPLE).

16. Recall that a considerable degree of correctness and consistency is already ensured by the inheritance architecture of the lexicon and by the lexicographer's workbench LIA.

For readings with complements, the examples contain an annotation of the syntactic complement structure (which is automatically provided by the workbench LIA). After the lexicon has grown by a defined amount, all lexicon examples are parsed by the WOCADI parser, which reports the following error types: The sentence cannot be parsed as a whole or the parse quality is suboptimal (e.g. due to unattached complements) or the parse does not contain the corresponding concept ID (but the ID of a homonymous or polysemous reading). With the help of these messages, the lexicographer is able to trace and correct lexicon errors.

Empirical validation. HaGenLex is constantly empirically validated by parsing corpora. These corpora consist mainly of newspaper articles and scientific abstracts. Currently, they comprise over 170 million words. Since these corpora are not annotated with parse trees or similar information, validation by parsing these corpora usually does not allow to trace errors back to the involved lexemes, but only serves as an indication of possible lexeme errors. Despite its heuristic nature, this method of spotting errors in the lexicon has proved to be worthwhile.

We are working with the following heuristics, besides others: All sentences of a corpus are parsed by WOCADI and the relative frequency f of successful parses is determined. For each lexeme l in the lexicon, the expected relative frequency of a successful parse is assumed to be equal to f if one looks only at corpus sentences containing l. (If there are several polysemous readings, one can presume an equal distribution across these readings or use reading frequencies derived from other sources to modify the expected relative frequency.) Then the actual relative frequency of lexeme l in the parse results for the corpus is calculated. If this actual frequency is significantly lower (or higher) than the expected frequency, a warning is generated that a lexicographer should have a look at this lexeme (*unexpected frequency heuristic*). To avoid repeated false warnings, one can accept this unexpected frequency by a special lexeme annotation. A low frequency is most often caused by (partially) incorrect specifications of syntactic or semantic features, e.g. too narrow selectional restrictions for verbs, nouns, or adjectives; incorrect gender for nouns; errors in the morphologic paradigm assignment derived from CELEX and a manually maintained source; etc.

A second approach is the *high ambiguity rate heuristic*. In case the parser, which is quite specialized and successful for disambiguation tasks, cannot disambiguate between polysemous readings in many cases, a warning is shown. After a considerable part of the corpus has been reparsed with the improved lexicon, the described error spotting process is repeated with a set of selected heuristics and selected settings of parameters for these heuristics.

Use in parser-based applications. The WOCADI parser – and thereby all applications built on top of WOCADI – use HaGenLex as their main lexical knowledge source. These applications comprise natural language interfaces to different databases: bibliographic databases, meteorological databases containing multimedia documents, product catalogs, etc. The users of these applications provide reports of errors, anomalies, etc. that often lead to improvements of HaGenLex.

Checking against other sources. HaGenLex is continually validated relative to other lexical resources. In particular, the linking of HaGenLex entries to GermaNet entries is employed to check the consistency of the HaGenLex projections of GermaNet synsets. For instance, if two presumably synonymous HaGenLex entries are incompatible with respect to their semantic sort, then either the GermaNet synset is defective or the HaGenLex-GermaNet linking is wrong or there is an inconsistency in HaGenLex. The latter two cases give rise to modifications of HaGenLex entries.

Coverage guidance by corpus word form lists. Besides the correctness of entries, the quantity and relevance of covered word form tokens in real world texts is a second important quality of a lexical resource. To improve this quality, a frequency list of word forms that are currently unknown to the parser (using HaGenLex, compound analysis, name lexica, complex named entity parsers, etc.) is maintained. The most frequent new entries are manually tagged with a problem reason from a predefined inventory and if this reason is a missing lexical entry, this entry is added to the lexicon. Similarly, word forms that are covered only by entries in the underspecified lexicon derived from CELEX (which contains morpho-syntactic information) can be collected and these entries are extended to full HaGenLex entries by a completing run of LIA.

7. Conclusion and Prospects

The lexicon conception of HaGenLex realizes both an application independent lexical theory that makes use of the MultiNet semantic representation formalism and a knowledge base for syntactico-semantic analysis. Central features of HaGenLex are its emphasis on semantic aspects and its technological embedding, which is supported by the consistent application of the interoperability criterion of MultiNet. HaGenLex has proven to be of sufficient size and quality for real-world NLP applications like natural language interfaces to databases and information retrieval.

HaGenLex has reached a stage of development where it can be successfully employed as a basis for an iterative bootstrapping process to acquire additional lexical information semi-automatically. The idea is that at each step of the bootstrapping process new hypotheses for the lexical characterization of unknown words or readings are automatically generated by running the WOCADI parser over large text corpora on the basis of the current lexicon. Such a run generates linguistic contexts that serve as input for machine learning tools which deliver syntactico-semantic features of unknown words or readings. The updated lexicon undergoes a validation process by using it in applications and by systematic cross-checking supported by corpus and lexicon tools. In some cases, lexemes must be intellectually validated and manually edited by means of the lexicographer's workbench LIA. Increasing the quality and quantity of HaGenLex entries leads to better hypotheses and thus to less post-processing.

8. References

[BAA 95] BAAYEN R. H., PIEPENBROCK R., GULIKERS L., *The CELEX Lexical Database. Release 2 (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, 1995.

[BRA 85] BRACHMAN R. J., SCHMOLZE J. G., "An Overview of the KL-ONE Knowledge Representation System", *Cognitive Science*, vol. 9, num. 2, 1985, p. 171–216.

[BRE 01] BRESNAN J. W., *Lexical-Functional Syntax*, Blackwell, Oxford, England, 2001.

[BRI 93] BRISCOE T., COPESTAKE A., DE PAIVA V., Eds., *Inheritance, Defaults, and the Lexicon*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, England, 1993.

[BUI 98] BUITELAAR P., "CoreLex: An Ontology of Systematic Polysemous Classes", GUARINO N., Ed., *Formal Ontology in Information Systems. Proceedings of FOIS'98*, IOS Press, Amsterdam, The Netherlands, 1998.

[CAR 92] CARPENTER B., *The Logic of Typed Feature Structures*, Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, New York, 1992.

[COP 93] COPESTAKE A., SANFILIPPO A., BRISCOE T., DE PAIVA V., "The ACQUILEX LKB: An Introduction", Briscoe et al. [BRI 93], p. 148–163.

[DRO 95] DROSDOWSKI G., Ed., *Duden*, vol. 4, Dudenverlag, Mannheim, Germany, 5 edition, 1995.

[GNÖ 02] GNÖRLICH C., "Technologische Grundlagen der Wissensverwaltung für die automatische Sprachverarbeitung", PhD thesis, FernUniversität Hagen, Fachbereich Informatik, Hagen, Germany, 2002.

[HAR 94] HARTRUMPF S., "IBL: An Inheritance-Based Lexicon Formalism", AI-report num. 1994-05, 1994, University of Georgia, Artificial Intelligence Center, Athens, Georgia.

[HAR 96] HARTRUMPF S., "Redundanzarme Lexika durch Vererbung", Master's thesis, Universität Koblenz-Landau, Koblenz, Germany, June 1996.

[HAR 99] HARTRUMPF S., "Hybrid Disambiguation of Prepositional Phrase Attachment and Interpretation", *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, College Park, Maryland, 1999, p. 111–120.

[HAR 02] HARTRUMPF S., HELBIG H., "The Generation and Use of Layer Information in Multilayered Extended Semantic Networks", SOJKA P., KOPEČEK I., PALA K., Eds., *Proceedings of the 5th International Conference on Text, Speech and Dialogue (TSD 2002)*, Lecture Notes in Artificial Intelligence LNCS/LNAI 2448, Brno, Czech Republic, Sep. 2002, p. 89–98.

[HAR 03] HARTRUMPF S., *Hybrid Disambiguation in Natural Language Analysis*, Der Andere Verlag, Osnabrück, Germany, 2003.

[HEL 97] HELBIG H., HARTRUMPF S., "Word Class Functions for Syntactic-Semantic Analysis", *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria, Sep. 1997, p. 312–317.

[HEL 00] HELBIG H., GNÖRLICH C., LEVELING J., "Natürlichsprachlicher Zugang zu Informationsanbietern im Internet und zu lokalen Datenbanken", SCHMITZ K.-D., Ed., *Sprachtechnologie für eine dynamische Wirtschaft im Medienzeitalter*, Wien, Austria, 2000,

TermNet, p. 79–94.

[HEL 01] HELBIG H., *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*, Springer, Berlin, 2001.

[HEL 02] HELBIG H., GNÖRLICH C., "Multilayered Extended Semantic Networks as a Language for Meaning Representation in NLP Systems", GELBUKH A., Ed., *Computational Linguistics and Intelligent Text Processing*, vol. 2276 of *LNCS*, Berlin, 2002, Springer, p. 69–85.

[KUN 01] KUNZE C., WAGNER A., "Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche", LEMBERG I., SCHRÖDER B., STORRER A., Eds., *Chancen und Perspektiven computergestützter Lexikographie*, vol. 107 of *Lexicographica Series Maior*, p. 229–246, Niemeyer, Tübingen, Germany, 2001.

[LEV 93] LEVIN B., *English Verb Classes and Alternations. A Preliminary Investigation*, University of Chicago Press, Chicago, 1993.

[LEV 02] LEVELING J., HELBIG H., "A Robust Natural Language Interface for Access to Bibliographic Databases", CALLAOS N., MARGENSTERN M., SANCHEZ B., Eds., *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002)*, vol. XI, Orlando, Florida, Jul. 2002, International Institute of Informatics and Systemics (IIIS), p. 133–138.

[POL 94] POLLARD C., SAG I. A., *Head-Driven Phrase Structure Grammar*, Studies in Contemporary Linguistics, University of Chicago Press, Chicago, Illinois, 1994.

[PUS 95] PUSTEJOVSKY J., *The Generative Lexicon*, MIT Press, Cambridge, Massachusetts, 1995.

[QUI 68] QUILLIAN M. R., "Semantic Memory", MINSKY M., Ed., *Semantic Information Processing*, p. 227–270, MIT Press, Cambridge, Massachusetts, 1968.

[RUS 92] RUSSELL G., BALLIM A., CARROLL J., WARWICK-ARMSTRONG S., "A Practical Approach to Multiple Default Inheritance for Unifi cation Based Lexicons", *Computational Linguistics*, vol. 18, num. 3, 1992, p. 311–337.

[SAN 93] SANFILIPPO A., "LKB Encoding of Lexical Knowledge", Briscoe et al. [BRI 93], p. 190–222.

[SCH 98] SCHULZ M., "Eine Werkbank zur interaktiven Erstellung semantikbasierter Computerlexika", PhD thesis, FernUniversität Hagen, Fachbereich Informatik, Hagen, Germany, Aug. 1998.

[STA 02] STAUDT D., "Automatische Akquisition lexikalisch-semantischer Informationen aus Textkorpora: Eine Untersuchung zur Kategorie der Nomina", Magisterarbeit, Ruhr-Universität Bochum, Bochum, Germany, Aug. 2002.

[WAH 00] WAHLSTER W., Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin, Germany, 2000.

A. Selected Representational Elements of MultiNet

Name	Meaning	Examples	
		+	−
ANIMAL	animal	*fox*	*person*
ANIMATE	living being	*tree*	*stone*
ARTIF	artifact	*house*	*tree*
AXIAL	object having a distinguished axis	*pencil*	*sphere*
GEOGR	geographical object	*the Alps*	*table*
HUMAN	human being	*woman*	*ape*
INFO	(carrier of) information	*book*	*grass*
INSTIT	institution	*UNO*	*apple*
INSTRU	instrument	*hammer*	*mountain*
LEGPER	juridical or natural person	*fi rm*	*animal*
MENTAL	mental object or situation	*pleasure*	*length*
METHOD	method	*procedure*	*book*
MOVABLE	object being movable	*car*	*forest*
POTAG	potential agent	*motor*	*poster*
SPATIAL	object having spatial extension	*table*	*idea*
THCONC	theoretical concept	*mathematics*	*pleasure*

Table 1. *Semantic features ([SCH 98])*

Relation	Signature	Short description
AFF	$[si \cup abs] \times [si \cup o]$	Affected object
AGT	$[si \cup abs] \times o$	Agent
AVRT	$[dy \cup ad] \times o$	Averting/Turning away from an object
BENF	$[si \cup abs] \times [o \setminus abs]$	Benefactee
EXP	$[si \cup abs] \times o$	Experiencer
INIT	$[dy \cup ad] \times [si \cup o]$	Relation specifying an initial state
INSTR	$[si \cup abs] \times co$	Instrument
MCONT	$[si \cup o] \times [si \cup o]$	Mental content
METH	$[si \cup abs] \times [dy \cup ad \cup io]$	Method
MEXP	$[si \cup abs] \times d$	Mental carrier of a state
OBJ	$[si \cup o] \times [si \cup o]$	Neutral object
OPPOS	$[si \cup o] \times [si \cup o]$	Entity being opposed by a situation
ORNT	$[si \cup abs] \times o$	Orientation towards something
RSLT	$[si \cup abs] \times [si \cup o]$	Result
SCAR	$[st \cup as] \times o$	Carrier of state
SSPE	$[st \cup as] \times ent$	State specifi er

Table 2. *Cognitive roles*

entity [*ent*]	
object [*o*]	
concrete object [*co*]	
discrete object [*d*]	*house, apple, tiger*
substance [*s*]	*milk, honey, iron*
abstract object [*ab*]	
attribute [*at*]	
measurable attribute [*oa*]	*height, weight, length*
non-measurable attribute [*na*]	*form, trait, charm*
relationship [*re*]	*causality, similarity, synonymy*
ideal object [*io*]	*religion, justice, criterion, category*
abstract temporal object [*ta*]	*Renaissance, Easter, holiday*
modality [*mo*]	*necessity, intention, permission*
situational object [*abs*]	
dynamic situational object [*ad*]	*race, robbery, movement*
static situational object [*as*]	*equilibrium, sleep*
situation [*si*]	
dynamic situation [*dy*]	
action [*da*]	*write, sing, sell, drive*
happening [*dn*]	*rain, decay, explode*
static situation [*st*]	*stand, be ill*
situational descriptor [*sd*]	
time [*t*]	*yesterday, Monday, tomorrow*
location [*l*]	*here, there*
modal situational descriptor [*md*]	*impossible, necessary, desirable*
quality [*ql*]	
property [*p*]	
total quality [*tq*]	*dead, empty, green*
gradable quality [*gq*]	
measurable quality [*mq*]	*small, expensive*
non-measurable quality [*nq*]	*friendly, tired*
relational quality [*rq*]	*inverse, equivalent, similar*
functional quality [*fq*]	
operational quality [*oq*]	*fourth, last, next*
associative quality [*aq*]	*chemical, philosophical*
quantity [*qn*]	
quantifi cator [*qf*]	
numerical quantifi cator [*nu*]	*one, two, fi ve, hundred*
non-numerical quantifi cator [*nn*]	*all, many, several*
unit of measurement [*me*]	*kg, meter, mile*
measurement [*m*]	*three miles, two hours*
graduator [*gr*]	
qualitative graduator [*lg*]	*very, especially, rather*
quantitative graduator [*ng*]	*almost, nearly, approximately*
formal entity [*fe*]	(meta level entities like fi gures and tables)

Table 3. *Hierarchy of ontological sorts*

Name	Meaning	Examples
FACT	The *facticity* of an entity, i.e. whether it is really existing (value *real*), not existing (value *nonreal*), or only hypothetically assumed (value *hypo*).	*"IBM"* [FACT *real*], *"UFO"* [FACT *nonreal*], *"String"* [FACT *hypo*] (*"String"* as the reading from astrophysics).
GENER	The *degree of generality* indicates whether a conceptual entity is generic (value *ge*) or specific (value *sp*).	*"admiral"* [GENER *ge*], *"Nelson"* [GENER *sp*].
QUANT	The intensional aspect of *quantification* specifies whether the concept is a singleton (value *one*) or a multitude (value *mult*) with the subtypes *fquant* and *nfquant* for fuzzy and non-fuzzy quantifiers, respectively.	*"a"* [QUANT *one*], *"four"* [QUANT *mult*], *"many"* [QUANT *fquant*], *"all"* [QUANT *nfquant*].
REFER	The *determination of reference*, i.e. whether the concept determines the reference (value *det*) or not (value *indet*). This type of characteristic plays an important part in text processing, especially for reference resolution.	*"this"* [REFER *det*], *"a"* [REFER *indet*].
CARD	The *cardinality* characterizes the extensional aspect of a multitude. Such cardinalities are useful, among others, for the disambiguation of coreferences. In the lexicon, they are only used in connection with numerals transferring their cardinality to the representative of a noun phrase (when used in such a combination).	*"three"* [CARD 3], *"five"* [CARD 5], *"several"* [CARD > 2].
ETYPE	The *type of extensionality* of an entity with values: *nil* – no extension, 0 – individual which is no set, 1 – entity with a set of elements from type [ETYPE 0] as extension, etc.	*"Napoleon"* [ETYPE 0], *"crew"* [ETYPE 1].
VARIA	The *variability* describes whether an object is conceptually varying (value *var*) or not (value *con*).	In *"This teacher likes every student"*: *"This teacher"* [VARIA *con*], *"every student"* [VARIA *var*].

Table 4. *Layer attributes (the first four describe intensional aspects, the remaining three describe extensional aspects)*