Reducing Lexical Redundancy by Augmenting Conceptual Knowledge

Sven Hartrumpf and Marion Schulz

Applied Computer Science VII, University of Hagen, 58084 Hagen, Germany
{Sven.Hartrumpf,Marion.Schulz}@FernUni-Hagen.De

1 Systematic Ambiguity

Lexical (and structural) ambiguities make language as expressive as it is. Computational lexicons thus have to cope with a large amount of polysemous words. Research in the last decade (e. g., Pustejovsky (1995), Kilgarriff and Gazdar (1995)) has aimed at identifying different types of polysemy in order to capture underlying regularities. This paper deals with a subtype of regular polysemy illustrated by the following sentences:

(1) *Mrs. Richwoman donated a significant amount to the opera* (institution).
(2) *The opera* (building) *was built by a famous architect.*
(3) *The opera* (staff) *was on strike yesterday.*
(4) *The opera* (contents) *features a Russian Tsar.*
(5) *I enjoyed the opera* (performance) *very much.*
(6) *I bought the opera* (recording) *at WOM's.*

As you can see the English word *opera* stands for a range of concepts which are closely interrelated. Bierwisch (1983) introduces the name *Konzeptfamilie* (**concept family**) for this range of possible interpretations of lexemes denoting institutions or cultural performances.[1] We call this kind of polysemy **inherent polysemy** because a single occurrence of a lexeme of this special subtype of regular polysemy can simultaneously be interpreted as different **variants**[2]. A variant then, in contradistinction to a reading of other types of polysemes, need not be the exclusive interpretation of the polysemous word:

(7) *I enjoyed very much the opera* (performance/contents) *featuring a Tsar yesterday in the local opera house.*

A member of the concept family may be lexicalized as a separate lexeme, like the building variant of the polyseme *city council*: it is lexicalized as *town hall*. This seems to imply that the members of the family have conceptual status of their own.

The aim of this paper, however, is not to postulate (cognitively adequate) concepts. We have a rather practical goal in view: the construction of a lexicon[3] where knowledge is located at (and inherited from) the most general place in the hierarchy possible. We group concepts into a concept family whenever this family is lexicalized by an inherently polysemous lexeme. In order to minimize redundancy and maximize reuse of information, especially when building language lexicons for several languages, one should maximize the information stored in the concept lexicon and thereby minimize the information in the language lexicon.

Although concept lexicon and language lexicons are distinguished, both are unified in one (inheritance) structure. A combined treatment has been advocated for several reasons:

[1] Bierwisch mentions additional interpretations for *Oper* without claiming completeness.

[2] Cruse (1995) uses the term *cooperative readings* (in contrast to *antagonistic readings*).

[3] The lexicon reported here is applied for example in the NatLink subproject, which provides a natural language interface to large distributed multimedia databases in the project *Virtual Knowledge Factory* (*Virtuelle Wissensfabrik*) of the state Nordrhein-Westfalen.

no duplication of information (Cunningham and Veale, 1991, p. 987)
cognitive adequacy (Dahlgren, 1988, p. 18) argues for identifying word meaning with concept. This suggests storing concepts and lexical entries in one structure.
linguistic adequacy Linguistic and nonlinguistic knowledge should be represented on the same level (cf. Jackendoff (1983)).
In addition to these advantages, we see three other advantages:
small redundancy Combined lexicons allow to isolate information in the concept lexicon thereby removing redundancy from several language lexicons.
reuse of knowledge One can reuse the knowledge encoded in a concept lexicon across different grammar theories and across different languages.
relevance for machine translation The concepts defined in the concept lexicon can be thought of as an interlingua.

2 Lexicon Formalism

We use the lexicon formalism **IBL** (**inheritance-based lexicon formalism**), which Hartrumpf (1996) has developed and implemented. IBL is heavily inspired by the lexicon formalism of the *Environnement Linguistique d'Unification* (ELU), which uses multiple default inheritance as described by Russell et al. (1992). IBL differs from ELU's formalism in four main ways: feature structures may contain complex disjunctions and complex negations; predicative constraints can use coroutining to wait for arguments to become sufficiently instantiated; a class can decide where in the feature structure to inherit information from a superclass (**locating inheritance**); IBL is strongly typed, i. e., all feature structures must have a type.

An IBL lexicon consists of type, generator and class definitions. Types are defined to type feature structures. A generator is a rule that can be used to generate a new class from another class. The most important concept are classes. A class is defined by its name, a list of direct superclasses (superclass list), a (possibly empty) main feature structure containing definite information, a (possibly empty) default feature structure containing default information, and a set of variant feature structures defining a set of mutually exclusive alternatives.

The default information coming from a superclass can be overwritten in subclasses. Inheritance conflicts are solved by prioritized inheritance which uses the concept of class precedence lists (CPLs) from the Common Lisp Object System. The information of a superclass can be inherited by a subclass at a certain position (locating inheritance); such a position can be a feature path or an element of a set value, list value or a disjunction. The information inherited cannot be restricted or selected because this contradicts the principle of data encapsulation and may lead to unwieldy and poorly structured lexicons. To calculate the extension of a lexical class (a class describing a single lexeme), one adds to it the information from main/variant feature structures of its superclasses by unification and the information from default feature structures of its superclasses by default unification.

3 Combining Concept Lexicon and Language Lexicon(s)

We distinguish the **concept lexicon** and language-specific lexicons (**language lexicons**) and call the combination of a concept lexicon with a language lexicon a **combined lexicon**. Both parts use the inheritance formalism described in section 2. This allows a seamless integration of conceptual and linguistic knowledge. To achieve this integration, lexemes inherit from concepts all the information that is language-independent.

A class describing a single lexeme adds the language-specific information by inheriting from classes higher in the language lexicon hierarchy and by stating the lexeme's idiosyncratic properties like the base form.

The basic categories for describing concepts are taken from the knowledge representation language MESNET (Multi-layered Extended Semantic NETworks, cf. Helbig and Schulz (1997)). A concept is described by an IBL class containing a feature structure of type *concept*. It specifies the concept's semantics, information about semantic argument structure and further compatible semantic relations with the help of MESNET's classificatory means. Using IBL's multiple inheritance with defaults, the concept lexicon is built as an inheritance hierarchy of classes.

A lexeme is described by an HPSG-like feature structure. Morphological (feature MORPH), syntactic (feature SYN), and semantic/subcategorization information (feature SEMSEL) is distinguished. The latter, which is central for the purpose of our paper, is a set-valued feature whose elements have type *semsel*. This type is a subtype of *concept* as a lexeme inherits semantic information from concept classes and might have additional semantic features under the feature SEMSEL. This feature has sets as values in order to allow all readings of a polyseme to be stored with small redundancy; common information under the features MORPH and SYN is shared and only one feature structure is needed to represent a polyseme.

Due to the thorough definition of types for concepts and features, combing concepts and lexemes is straightforward: The concept lexicon is an inheritance hierarchy defining concept classes by feature structures of type *concept*; the language lexicon is an inheritance hierarchy which embeds by locating inheritance information from the concept lexicon as elements of the set-valued feature SEMSEL.

4 Inherent Polysemy in Combined Lexicons

The definition of a concept is a value of type concept: for a concept family, it is a disjunctive one; for all other concepts, a nondisjunctive one. In the language lexicon, an "ordinary" polyseme (one not allowing switching between readings) has more than one element in the set value for feature SEMSEL (Each element can be a disjunctive value introduced for a concept family on the conceptual level.); all other lexemes have just one element under SEMSEL.

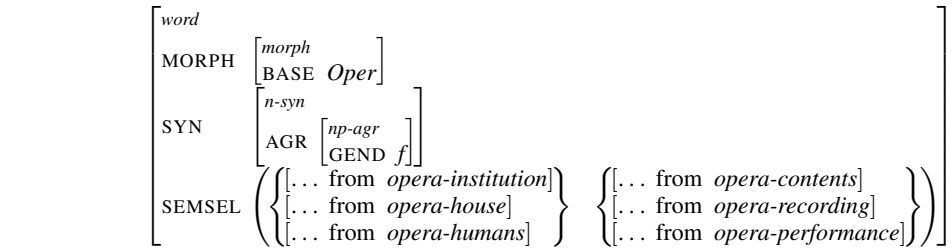
Bierwisch (1983) distinguishes two *Unterfamilien* (subfamilies) for the German noun *Oper*. They relate to a subconcept of institution-f and a subconcept of performance-f, respectively. In our opinion, both subconcepts constitute independent concepts that happen to be lexicalized by the same lexeme in German (and some other languages). This lexeme (an "ordinary" polyseme) has two readings (corresponding to the concepts *opera-institution-f* and *opera-performance-f*) which are an inherent polyseme each.

The concept opera-performance-f is defined by locating inheritance forming a disjunctive value with three disjunction elements inherited from opera-contents, opera-recording, and opera-performance (similarly for opera-institution-f). In addition, the concept opera-performance-f is defined by the superconcept performance-f, which has three more general variants, and is refined by locating inheritance from those more specific variants mentioned above, which are all subconcepts of the corresponding variants of performance-f.

A lexeme of a language is described by its idiosyncratic properties (base form etc.) and is linked to a concept by inheriting from the corresponding concept class in the concept lexicon hierarchy.

Figure 1 shows the basic structure of the lexicon information for the German noun *Oper* which is derived by inheritance. The feature SEMSEL has two set elements, each consisting of three disjunction elements (only their origin, not their content is shown). The advantage of small redundancy becomes evident if one compares the feature structure in figure 1 (which is much larger if fully shown) with the three-line definition of the lexicon entry for *Oper*:

```
word "Oper"
  inherit ((opera-institution-f (semsel 1)) (opera-performance-f (semsel 2)) n-f)
  main ((morph base) "Oper"
```



Note: Curly braces indicate disjunctions, parentheses group set elements.

Fig. 1: Basic Structure of the Lexicon Information for *Oper* Derived by Inheritance

References

Bierwisch, Manfred (1983). Semantische und konzeptuelle Repräsentation lexikalischer Einheiten. In *Untersuchungen zur Semantik* (Edited by Ružička, Rudolf; Motsch, Wolfgang), studia grammatica XXII, pp. 61–99. Akademie-Verlag, Berlin.

Cruse, D. A. (1995). Polysemy and related phenomena from a cognitive linguistic point of view. In *Computational Lexical Semantics* (Edited by Saint-Dizier, Patrick; Viegas, Evelyne), pp. 33–49. Cambridge University Press, Cambridge, England.

Cunningham, Pádraig; Veale, Tony (1991). Organizational issues arising from the integration of the lexicon and concept network in a text understanding system. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, volume 2, pp. 986–991.

Dahlgren, Kathleen (1988). *Naive Semantics for Natural Language Understanding*. Kluwer Academic Publishers, Boston.

Hartrumpf, Sven (1996). *Redundanzarme Lexika durch Vererbung*. Master's thesis, Universität Koblenz-Landau, Koblenz.

Helbig, Hermann; Schulz, Marion (1997). Knowledge representation with MESNET: A multilayered extended semantic network. In *Working Notes of the 1997 AAAI Spring Symposium on Ontological Engineering*, pp. 64–72.

Jackendoff, Ray (1983). *Semantics and Cognition*. MIT Press, Cambridge, Massachusetts.

Kilgarriff, Adam; Gazdar, Gerald (1995). Polysemous relations. In *Grammar and Meaning: Essays in Honour of Sir John Lyons* (Edited by Palmer, F. R.), pp. 1–25. Cambridge University Press, Cambridge, England.

Pustejovsky, James (1995). *The Generative Lexicon*. Number 13 in CSLI Lecture Notes. MIT Press, Cambridge, Massachusetts.

Russell, Graham; Ballim, Afzal; Carroll, John; Warwick-Armstrong, Susan (1992). A practical approach to multiple default inheritance for unification based lexicons. *Computational Linguistics*, 18(3):311–337.